Stockholm School of Economics
Department of Economics
5350 Master's thesis in economics
Academic year 2021-2022

When faced with danger, seek refuge in the herd: A study on the stability of cooperation strategies under evolving incentives.

Ethan O'Leary (41643)

Abstract

Understanding the mechanisms that drive human non-kinship cooperation is an important aim of behavioural economic literature. However, many results are based on observations in abstract games under context-specific conditions which limits the external validity of conclusions. With time, these abstract games have become more complex and have relaxed their restrictive assumptions, but are still far from modelling the intricacies of human decision making in society. One such limitation of cooperation experiments that has received little attention is the assumption that individuals formulate constant strategies that may or may not be conditional on the social norm. This thesis studies the dynamics of cooperative phenotypes under varying levels of high external threat using a series of fixed effects linear probability models applied to data sourced from the US TV series: *Survivor*. I find that as the external threat increases, individuals display a decreasing propensity to cooperate and a decreasing sensitivity to the social norm for cooperation. A non-linear analysis reveals the interdependent relationship between social norms and private payoff incentives. Hence, I propose that to maintain cooperation in high-threat settings, it must be that either there is a strong social norm for cooperation that includes credible social punishments, or that the threat is framed as surmountable, which may then endogenously create such a social norm.

Keywords: Cooperation, Social norms, Strategy, Field experiment, Risk.
JEL:C70, C73, C93, D91

Supervisor: Magnus Johannesson
Date submitted: 05-12-2021
Date examined: 17-12-2021
Discussant: Axel Granström
Examiner: Kelly Ragan

Contents

List of Tables

Acknowledgements

I'd like to express my utmost gratitude to a number of individuals who were integral in the process of writing this thesis. First, I'd like to thank my supervisor, Magnus Johannesson. This thesis was inspired by his course on Behavioural Economics and I am grateful that he agreed to supervise this project. Further, I thank Magnus for donating his time and energy to helping this thesis progress through times of difficulty. Without this inspiration or support, I would not be in this position.

Next, I would like to thank Jesper Roine and Anna Dreber for providing feedback to my research proposal on their own time. I further thank Jesper for his continued support and academic guidance in this last year.

Finally, I thank my partner Paula for the personal support throughout the project. I hope that this work is just the start of my research career and that it is the beginning of my contribution to what we believe in.

1 Introduction

The paradigm of studying human cooperation has shifted since the dawn of modern experimental economics. At the fundamental level, economists are handed the normative axioms of homoeconomicus. This sets the scene for a rational selfish decision making body interacting with its environment in search of perpetual personal gain. Under this light, economists observing cooperation may be bemused as to why subjects engage in seemingly costly but pro-social activities such as punishment or altruistic giving. As these tendencies have been identified in various contexts, the economic model of cooperation has evolved to incorporate fairness (Rabin, 1993), social preferences (Fehr and Schmidt, 1999), identity theory (Tajfel et al., 1971), and conformity (Bernheim, 1994).

With each observed anomaly, one may wonder how long these axioms can be moulded until they converge to a biological limit. Anthropologists support the theory that humans evolved to be instinctively highly cooperative creatures and this cultural norm was maintained through the threat of punishment in the form of ostracism or even death (Balikci, 1970; Lee, 1979).[1] Ironically, many modern problems that the global population now faces such as climate change and inequality, are rooted in the dearth of this once ubiquitous cooperation (Helénsdotter, 2019). Hence, what once was our natural status quo has now become a goal of social policy. Such disparity has birthed economic research on the positive axioms of cooperation. In this light, the primordial question of economic studies around cooperation has moved from ascertaining why humans cooperate towards understanding how policymakers can encourage societies to cooperate.

Most economic studies on cooperation focus on an abstract class of game termed social dilemmas, which define a choice between private individual incentives and social preferences (Olson, 1971). In these games, individuals have an incentive to cheat their partner and obtain some bonus for doing so, only cooperating when the threat of punishment is large enough or the future looms heavy (Dawes and Thaler, 1988). This means that the instantaneous Nash-equilibrium response of each individual is to free-ride and reap benefits realised by the actions of other cooperators. However, the total surplus and welfare of the group is maximised when all members cooperate. This class of game is extremely generalisable but is often studied in strict laboratory settings under known monetary payoff matrices, anonymous interaction, and restricted time frames. While these features help researchers claim causal relationships, the common setup achieves very little when one wishes to apply findings to field decisions and policies where environments are not controlled.

On observing behaviour in these games, researchers have been able to bridge the gap between homoeconomicus and the cooperative primate. Initial experiments attempted to divide individuals into purely selfish free-riders and those who showed degrees of cooperation. This dichotomy was greatly supported through the seminal work of Axelrod and Hamilton (1981), who showed that cooperative tendencies may in fact be strategically instrumental and can increase one's long-run income, and as such, cooperators need not be selfless. The authors designed a study in which game theorists contributed strategies to a computerised game. They found that a conditionally cooperative matching strategy termed 'Tit-for-Tat' was dominant and robust to invasions of rival strategies. This strategy involved cooperating on one's first decision and then matching the previous decision of one's co-player ad infinitum. Further research attempted to distinguish between discrete types of this coined conditional cooperation to identify the underlying mechanism which motivates the emergence and persistence of such a game plan.

[1] This theory is formalised in the self-domestication hypothesis. See Bednarik (2008) for a well versed summary.

One such mechanism, the theory of reciprocity, was born out of the finding that individuals interact with higher rates of cooperation when there is sufficient potential for continued relationships (Murnighan and Roth, 1983). This theory dictates that individuals are inclined to meet kindness with kindness and evil with evil. Thus, an individual may be more inclined to cooperate if they are aware that their co-player will react reciprocally in the next period, since they may avoid negative punishment and may be positively rewarded. However, this mechanism is only relevant if there is a chance that there will be a next period to realise the outcome of this reciprocity.

Theories of reciprocation have extended the matching 'Tit-for-tat' strategy to incorporate various degrees of realistic qualities into the discrete categorisation of cooperativeness. These include the idea of forgiveness (Murnighan and Roth, 1983; Dreber et al., 2014; Dal Bó and Fréchette, 2011; Embrey et al., 2018; Proto et al., 2019)), altruism (Fischbacher et al., 2001; Rustagi et al., 2010), and social cohesion (Tajfel et al., 1971; Traxler and Spichtig, 2010). Nonetheless, in order to argue for the external validity of categorising individuals by their tendency to cooperate given the decisions of other individuals, one needs to assume some form of general stability in the distribution of these prescribed phenotypes.[2] Hence, recent progress in the literature has explored the evolving emergence of conditional cooperation by identifying two channels of strategic instability.

Firstly, different groups of individuals may respond to identical incentives with different behaviours depending on their cultural values, backgrounds, and ideologies. Bigoni et al. (2016) demonstrate that social groups have varying degrees of inherent preferences for cooperation by performing a variety of social dilemma games on different sub-populations in Italy. They show that even under identical control parameters and incentives, individuals can react very differently. Further, they show that the degree of cooperation can be very specific to social parameters and individual beliefs about the group-wide average value of such. Helénsdotter (2019) also shows that individuals who are more politically liberal are inclined to invest more in public goods than those of centrist and rightist ideologies. Moreover, experimenters cannot control how individuals interpret the information with which they are presented. The literature has established that individuals who experience cooperative relationships or specific social cues tend to apply these experiences to experimental decisions (Henrich et al., 2005). For example, Gneezy et al. (2016) study fishermen in Brazil who either work alone in lakes or in groups out at sea in a prisoner's dilemma game. They show that those who work in groups at sea are significantly more likely to cooperate in a controlled game and donate to social causes. This effect is identified even after controlling for self-selection in to communities of higher degrees of cooperation. These findings are however based on specific and stable game parameters and therefore do not show the dynamic strategy set that may be employed.

The second channel of instability proposes that different incentives may encourage varying degrees of strategic cooperation to emerge even within identical groups of individuals. Pillutla and Chen (1999) show that the definition of the dilemma itself can alter individual propensity to cooperate. They demonstrate that the more pro-social an action is implied to be, the more inclined people are to cooperate and individuals hold higher beliefs that others will match this action. Arechar et al. (2018) systematically vary the probability of game continuation between two extremes to show that individual strategy choice is a learnt concept specific to the game environment. The authors find that conditional cooperation is something that is learnt over time in long games, but defecting is the modal strategy in short games. Moreover, Nagatsu et al. (2018) provide an example of when

[2] I refer to an individual phenotype as their characteristic cooperation strategy in a game as according to Rossetti et al. (forthcoming). Thus a phenotype not only defines if an individual conditionally cooperates but also how lenient they are to defection and how sensitive they are to group decisions.

even the least cooperative players may be incentivised to cooperate by employing signals of social expectations among fellow non-cooperators after the removal of cooperators from the situation. Finally, Dreber et al. (2014) suggest that situational complexity may bring about greater defection, but that pro-social behavioural strategies are paramount to observe cooperation when it is not a private equilibrium action.

On identifying these mechanisms of strategy variation, one is able to incorporate deterministic and stochastic components into a continuous definition of one's propensity to conditionally cooperate (Traxler and Spichtig, 2010). In this thesis, I hence ask how one's incentive to conditionally cooperate, and thus one's cooperative *phenotype*, varies as the exogenous probability that one will interact again decreases while simultaneously measuring the observed social norm for cooperation within a group.

Specifically, I aim to contribute to the growing literature which tests the robustness of strategic categorisation of individuals by systematically varying the marginal return to cooperation and subsequently identifying the interaction of this variable with the behaviours of others to observe individual conditional strategy. To achieve this, I partition the marginal return to cooperation into two distinct factions. First, the private incentive to cooperate is defined as the benefit an individual receives from cooperation regardless of the number of fellow cooperators. To illustrate this, consider an individual who is deciding whether to recycle their milk carton (cooperate) or to throw it in the landfill (defect). The individual's private incentive to cooperate is their share of the environmental benefit of having that milk carton recycled instead of disposing of it in a landfill, minus the cost of walking to the recycling facility. The second component is the social incentive to cooperate, which is defined as the social status benefit an individual receives from behaving in accordance with a social norm. In the recycling example, this is the social recognition that one receives by adhering to a local norm of recycling or the social shunning one feels should they violate such a norm. Since social incentives are endogenously determined by integrated interactions, the purpose of this study is to identify how policymakers can influence the apparent private incentives, namely the manageability of an external threat, in order to maximise cooperation given some level of group-wide social norm. I aim to replicate the results of Dreber et al. (2014), who find that as private incentives to cooperate disappear, social incentives can still maintain some degree of conditional cooperation. I expand on their findings by attempting to account for this effect through estimating one's evolving sensitivity to the endogenous social norm (Traxler and Spichtig, 2010). Further, I aim to expand on the study of Arechar et al. (2018) who show that conditional cooperation is abandoned when one transitions to a game with a low probability of continuation from a game with a high probability of continuation. I do so by presenting the dynamic mechanism of a continuous adjustment from a high marginal return to cooperation to a low marginal return to cooperation. This examines the continuous evolution of one's propensity to conditionally cooperate in face of an evolving environment.

The literature is overwhelmingly inconclusive on its ability to present cohesive and externally valid conclusions on cooperation decisions. Observing specific individuals in specific environments under some monetised incentives cannot be representative of the multitudes of social dilemmas and may not be consistent for long periods in larger societies. This thesis departs from laboratory standards and studies cooperation in the field under very high payoffs and an evolving marginal return to cooperation. This is chosen to observe individuals outside of their social milieu under survival-like parameters in an effort to capture innate behaviours. Thus, instead of observing abstract and controlled games; relaxing assumptions one at a time, this study attempts to observe a situation close to one under which cooperation has been shown to be natural by anthropologists. While the

study design renders causal conclusions impossible, I aim to uncover the inconsistency in individual reactions to incentives under strategic circumstances. I also diverge from the established and commonly employed method of strategic categorisation: the *structural frequency estimation method* as proposed by Dal Bó and Fréchette (2011). Instead, I model a continuous response to the decisions of other players in n-person games using the intuition of Experienced Weighted Attraction with Norm Psychology (EWANP) (Camerer and Ho, 1999) in an application of the discrete choice with social norms model by Brock and Durlauf (2001).

This behavioural model is appropriate since it separates the individual utility into two deterministic components. The first element of the utility function consists of one's private common net benefit from an action. I herein refer to this as the private incentives for cooperation. The second deterministic element of the utility function incorporates conditional strategy and measures the utility one obtains by acting in accordance with the common social norm.

The use of the EWANP model allows the continuous measurement of one's reciprocation and sensitivity to the norm. For example, an always defector (an individual who never cooperates) would not care for the social norm and thus have a zero sensitivity weighting towards such a norm. Conversely, with any type of strategy dependent on the decisions of others such as the tit-for-tat strategy, one may have some non-zero sensitivity to this norm at some time point and thus will derive utility from the accordance of one's actions to the group-wide average. The inclusion of such a continuous parameter to measure reciprocity has two advantages. First, this permits the identification of a dynamic strategy: a strategy where an individual conditionally cooperates dependent on the beliefs that they hold regarding private and social incentives (Realpe-Gómez et al., 2019). Second, the continuous model allows the control of some degree of group tolerance. In dyadic games, it is easy to observe reciprocation since one individual is responding to another. In n-person games however, it is more challenging to define a strategic conditional strategy since some conditional cooperators may defect as soon as one co-player defects, while others may only defect once a majority does so (Kurokawa et al., 2018). I herein refer to the incentive to conform to social norms according to one's social sensitivity as one's social incentives.

As aforementioned, I study individual behaviour under very high payoffs in order to emulate survival-like play. Therefore, I apply the model to data collected from contestant decisions in the US TV series: *Survivor*. Economists have been interested in studying game show behaviour for some time since high-stake interactions can be observed and thus fundamental predictions from game theory can be tested (Post et al., 2008). The game show *Survivor* constitutes a series of strategic interactions between a localised community of individuals who each attempt to *survive* the current round. At the end of each round, participants must vote to eliminate a player who then departs with a cash prize that increases with the time one survives in the game. Thus surviving a round in the game increases one's expected payoff. A number of articles have been written using observations from the series (see Karlan (2017)), but large scale studies have not yet been performed, perhaps due to the complexity of the data and heterogeneity across seasons. The use of data from *Survivor* however, provides a number of advantages to this study. First, the prize is on average one million dollars with runners up claiming six figure sums (Reality Blurred, 2010). Hence, interactions are under very high stakes which minimises the risk of payoff related abstraction bias and permits the assumption that individuals interact purely according to the prize incentive rather than exhibiting non-payoff related altruistic preferences (Rabin, 1993). Second, the game show is unscripted, long and situational unlike laboratory experiments meaning that individuals live their daily lives in desolate locations under their own constructed shelters and are responsible for providing food for themselves. This further

limits the possibility of abstraction bias impacting a decision. Third, contestants are initially split into separate tribes or subgroups allowing the observation and definition of subsets of contestants and subsequent localised decisions. This also allows for the assumption of exogenous variation of private incentives. Fourth, the series has been running for more than forty seasons in the US alone with many international versions. Therefore, the potential pool of data on individual decisions in the game is very large and can be expanded from this study.

This thesis composes a panel data set from individual decisions three periods before and three periods after the specific 'merger' event where the sub-groups (herein referred to as neighbourhoods) join to form one super-tribe and interact together. Individual and game period fixed effects are then employed on a linear probability model to estimate the interaction of private incentives and social incentive parameters on the individual decision to cooperate. I define a cooperative behaviour as one in which a contestant is a member of a strategic alliance which is a set of rules stipulating that alliance members do not vote to eliminate each other and that they all follow a specified voting pattern. The identification of such alliances is performed using data from episode summaries, peer-reviewed fan-collected data and episode transcripts. Further, in order for a decision to be cooperative, one must also abide by the voting rule of the alliance. Contestants are free to form these alliances as they wish and are not asked if they intend to.

Private incentives are proxied by variations in the relative threat of elimination sourced from outside of one's neighbourhood: namely in the formation of rival alliances between members of other neighbourhoods. This measure is then weighted against the size of one's neighbourhood. The weighted measure ensures that a high value of this relative threat makes cooperation less attractive since the threat is less manageable (one's neighbourhood cannot overcome the threat even if they all cooperate) but still permits a small positive individual gain from cooperating regardless. A low value here does not indicate that there is no threat, but that the threat is manageable and that there will be a net positive gain from cooperating to overcome the opposition. Social incentives are proxied by the contemporaneous measure of the number of individuals in one's neighbourhood alliance who cooperate. I choose to use contemporaneous measures of social incentives instead of observed past actions since I observe a game in which perfect communication is available and thus individuals interact between decisions and signal their likelihood of cooperating through communication (Smerdon et al., 2016; Cooper and Kühn, 2016; Realpe-Gómez et al., 2019).

Specifically, I construct two fixed effect models. The first model studies the effect of cooperation on an individual's probability of survival controlling separately for the decisions of other neighbourhood members and the external threat. This model aims to uncover the private payoff of an individual's decision to cooperate and to observe if this marginal benefit of cooperation is jointly determined by social and private incentives.

The second model proceeds to investigate the interaction between private incentives and social incentives in the decision to individually cooperate. I present a theoretical model which predicts that individual propensity to cooperate is increasing in private and social incentives. Next, I propose that the two categories of incentives interact and can co-determine the emergence of a cooperative equilibrium. The model predicts that private incentives alone determine the emergence of a private equilibrium and social incentives prevent individuals from deviating from this. As such, I hypothesise that decreasing private incentives diminish the propensity of individuals to conditionally cooperate. I test this using an interaction term between the proxy for private incentives and the proxy for social incentives. If I can not reject this hypothesis then it is evidence that social strategies and thus coop-

eration phenotypes are not stable in between different marginal per capita returns to cooperation.

Finally, I test the sensitivity of this effect according to the similarity of salient social identities within neighbourhoods. It has been established in economic and ethnographic literature that there exists a significant in-group bias for cooperation (Tajfel et al., 1971; Henrich et al., 2005). However this bias is both variable between populations (Bigoni et al., 2016; Martinangeli and Martinsson, 2020; Drouvelis et al., 2021) and is variable depending on the fragility of the group definition (Kok et al., 2020). Helénsdotter (2019) suggests that this effect is driven by the beliefs of group members regarding whether co-players will cooperate. Accordingly, since homogeneous groups have similar individuals, individuals in these groups tend to have more trust in their co-players' competence as a fellow cooperator (Nagatsu et al., 2018). I employ a test of in-group bias in this study by including a set of variables which measure the similarity of neighbourhoods according to various salient identities.

The main results can be summarised as follows. First, falling private incentives to cooperate diminish the attractiveness of cooperation by crowding out the constant benefit from mutual cooperation. Second, positive private incentives are shown to increase one's probability of cooperation. Third, this effect does interact with the apparent social norm of a group such that decreasing private incentives decreases one's probability of adhering to a cooperative norm on average. Fourth, I show that the effect of private incentives is dependent on the strength of a social norm and whether the act of cooperation is incentive compatible (if there is a net private utility benefit from cooperating). Finally, I find no evidence for group similarity contingent effects.

The thesis continues with section 2 providing a discussion on the background to research of unstable conditional cooperation and on the existing literature on specific drivers of this. Section 3 describes the theoretical model that is applied to the study and the hypotheses to be tested. Section 4 then defines the data employed and the exclusion criteria that is necessary. Section 5 presents the identification method and empirical specifications along with translating my hypotheses to econometrically testable definitions. Section 6 presents the results to the regressions. Section 7 expands on the main experiments through a series of robustness checks. Section 8 discusses the results in light of the literature, the limitations of the study, and defines the results as a contribution to specific strands of the literature. Finally, section 9 concludes.

2 Literature Review

2.1 Background

At the rudimentary level, economics introduces the scholar to the notion of a rational and selfish agent: an individual in pursuit of personal gain while caring nothing for their peers. Researchers have tested this theory by employing social dilemma experiments such as the discrete prisoner's dilemma. In this game, individuals are faced with a choice between cooperation and defection such that it is in the collective interest to mutually cooperate, however each individual has a private incentive to defect and reap benefit from their opponent's cooperation. Under common rationality, the economic assumption for these games is that individuals acknowledge that each player has an incentive to deviate from the socially optimal action and thus cooperation breaks down. Nonetheless, it is not uncommon to observe rates of cooperation which range between 40-60% in such experimental games (Dawes and Thaler, 1988). Interestingly, as these economic games are repeated over time, the average rate of cooperation diminishes towards a known end of a relationship (Roth and Murnighan, 1978).

One explanation for these stylised facts was formalised by Axelrod and Hamilton (1981). The authors solicited a congregation of game theorists to submit their strategies to a tournament which matched pairs of players against each other for 200 iterations of the prisoner's dilemma game. The tournament found that the Nash Equilibrium of always playing defect was only stable in single games and as soon as the game was iterated, the dominant strategy was defined as dependent on a player's opponent's last move. This dominant strategy, termed Tit-for-Tat (TFT), describes a game plan such that the player commences with the decision to cooperate and thereafter copies their opponent's most recent realised action. TFT therefore rewards those who played cooperatively and punishes those who deviated through imitation. In other words, players eliciting a TFT strategy are reciprocal to their opponent and cooperate conditional on their opponent doing so. This study importantly provided the motivation for research into the mechanisms which drive such reciprocal cooperation, given that this strategy achieved the highest payoff in the tournament and has since been shown to be evolutionarily stable (Nowak and Sigmund, 1992).

The theory of reciprocity hence accommodates cooperative acts within fundamental economic theory since it prescribes a series of conditions under which cooperation maximises an individual's utility in the long run. One myopic explanation for the emergence of such behaviour may be derived from Adam Smith's (1759) conjecture that even a selfish agent may have some interest in the fate of others and as such some individuals may derive utility from other's utility. This theory of altruistic reciprocity was formally defined by Rabin's (1993) theoretical model of fairness such that individuals sacrifice their own material well-being in order to reward those who are nice or to punish those who are unjust, without seeking material gain or opportunistic leverage. Substantial evidence for this intrinsic reciprocity has been presented from experiments which isolate the motivation to punish without potential for future interaction (Fehr et al., 1993; Fehr and Gächter, 2000*b*; Fehr and List, 2003).

However, in an early study of cooperation in a repeated prisoner's dilemma game, Roth and Murnighan (1978) identified that individuals are only inclined to engage in cooperation when the probability that they will play another round is sufficiently high. The conclusion of this study is that individuals may be motivated by their potential for future benefits in their decision to cooperate. With this option for iterated interactions, Murnighan and Roth (1983) later conjectured that the emergence and success of reciprocation strategies may be due to rational and selfish forward-looking individuals. Dawes and Thaler (1988) refer to this strategic conditional cooperation as instrumental reciprocity which fundamentally differs from the innate altruistic reciprocity since it is presumed to be conditional on an environment and is implemented to leverage private reward.[3]

Many direct empirical studies of instrumental reciprocation have focused on the establishment of a punishment option and as such focus on negative instrumental reciprocity. Fehr and Gächter (2000*b*) employ a public goods game with two treatments: one was such that individuals could pay to punish other who has contributed insufficiently and one without such an option. They find that even when individuals interact with strangers, the option to punish significantly raises individual contributions to a public good. While it has been shown that individuals display some degree of altruism in anonymous interactions (Hoffman et al., 2000), Fehr and Gächter (2000*b*) argue that their results

[3]Throughout the empirical sections of this paper, I refer to conditional cooperation as reciprocal behaviour, however do not directly distinguish between intrinsic and instrumental behaviour. Under the high stakes described in the next section, I assume that I am only observing instrumental reciprocity however this assumption is not critical for my conclusions.

provide evidence that when faced with the risk of being punished, individuals are more likely to cooperate. Later, Dal Bó (2005) conducted a similar experiment but varied the information available to players regarding the certainty of game longevity. Although, the experiment was only performed over short-lived games, Dal Bó (2005) conducted two treatments of the repeated prisoner's dilemma game that had identical expected lengths, however one treatment granted participants certainty of the game's imminent finish point. The other treatment retained a degree of uncertainty such that participants were aware of a fixed probability of game continuation but not that they were playing a final iteration. The results of this experiment demonstrated that individuals cease to cooperate when they are aware of an imminent end to an interaction and thus future punishment can not be realised. Conversely, uncertainty of future interaction and potential punishment is enough to encourage continued cooperation.

This has led to the establishment of a necessary condition for cooperation to be stable under the theory of instrumental reciprocity in the form of a threshold strategy (Dal Bó, 2005; Blonski et al., 2011; Dal Bó and Fréchette, 2011). That is, a strategy which prescribes conditional cooperation when the weight of the future surpasses an empirically predictable critical value. This critical value is composed in relation to the game payoff parameters such that defection and the subsequent punishment that follows do not exceed the future contingent gains from mutual cooperation. Nonetheless, a critique of these theories is that threshold conditions describe abstract cases and require certain degrees of payoff and continuation probability transparency to be tested.

Paramount to the success of instrumental reciprocity in achieving a community-wide norm of conditional cooperation is that individuals formulate some belief that they will be rewarded for their pro-social actions in the future (e.g. Ackermann and Murphy (2019)) or that there exists a credible threat of future punishment (e.g. Deb and González-Díaz (2019)). The degree to which one punishes a violator or rewards a cooperator is however a topic for discussion. Strategic memory length, for example, describes one's patience or level of forgiveness (Murnighan and Roth, 1983). Empirical studies which compare the length of punishment periods have confirmed that individual strategies tend to be short in memory and as such forgiveness are common (e.g. Dreber et al. (2014)).[4] To justify these observations, Ali and Miller (2016) propose a theoretical argument that an unforgiving strategy may not be instrumental by definition if it results in extreme punishments such as ostracism or condemning an individual to permanent defection. They claim that this is because reciprocators may be better served by an attempt to return to a cooperation equilibrium through foregiveness. A generalised strategy was suggested by Breitmoser (2015) in which individuals return randomly to a cooperation phase after punishing a defector. Nonetheless, after studying for the presence of this strategy using four prisoner dilemma experiments, Breitmoser (2015) found that on average, individuals do seek to return to some cooperative equilibrium after only a brief period of punishment and that this forgiving strategy is more robust to environmental shocks and individual errors than strictly grim strategies. Hence, while forgiveness is important for robust equilibria, one may wonder about the danger of appearing to be too lenient if that leads to opportunistic defection.

Such reputation building strategies may serve the interest of social welfare if they lead to community-wide cooperation. However, individuals may use instrumental punishments to leverage a cooperative equilibrium in order to maximise their material gain when they eventually defect. Evidence for this opportunistic use of instrumental reciprocity was perhaps first identified by Fischbacher et al. (2001) who defined two types of reciprocal cooperators by soliciting individual contributions to a public

[4]Nonetheless, some intrinsic characteristics have been shown to increase one's degree of rationality such as intelligence and experience (Proto et al., 2019).

good at every level of communal contribution.[5] They find that while conditional cooperators are the most common type, some of these players tend to defect once the total contribution reaches a sufficient level. Thöni and Volk (2018) conducted a meta-analysis of replications of Fischbacher et al. (2001) using 7107 observed individuals and identified conditional cooperators as the modal phenotype at 61.3% of the sample, while 10.4% of the sample conditionally cooperated but only when the total social contribution was low.

While Kandori (1992) postulated that the mere presence of perfect conditional cooperators may lead to the community-wide breakdown of cooperation with a sufficient number of defectors, Reuben and Suetens (2011) demonstrated that unstable conditional cooperators such as those identified by Fischbacher et al. (2001) may also contribute to the realisation of inefficient cooperation breakdown if they wrongly expect others to continue contributing. Furthermore, Ackermann and Murphy (2019) added to this argument by showing that individuals who believe there to be a significant proportion of pro-social cooperators defect earlier in the game in anticipation that their actions may not be punished sufficiently. Thus individuals forming beliefs of other individuals' strategies may be able to get ahead of the game and defect to reap maximum reward when acting under forward looking reciprocation. Under this logic, if a cooperative norm is quickly established, the presence of opportunists can quickly unravel the efficient outcome which suggests that conditional cooperation may not be socially beneficial if it is not perfect in nature.

The above studies show that while the reciprocal phenotype is measurable, it may not be stable across environments nor across individuals. One explanation for this is that individuals are prone to stochastic errors and irrational choices. Nowak and Sigmund (1992), for instance, suggested that individuals may respond to errors by implementing some generous tit for tat strategy that assigns some probability to reciprocate. This stochastic modeling was formalised in Breitmoser's (2015) proposed semi-Grim equilibrium strategy which dictates that individuals follow a norm when groups mutually coordinate on a behaviour and randomise their decisions otherwise. However the assumption of stochastic strategies is contested for two reasons. First, from a modeling and policy perspective, reducing individuals to random decisions serves little purpose when one wishes to promote cooperation unless one can identify deterministic components of the decision. Further, humans have been shown to be poor randomisers in the field of anthropology (Gilovich et al., 2009).[6] While truly stochastic strategies may therefore be an unrealistic component, the incorporation of random errors and the beliefs of the distribution from which these errors are drawn is not nuanced in theory (Smerdon et al., 2016).

Nonetheless, since individuals acting in pure selfish interest may still elicit degrees of interdependent altruism (Sobel, 2005) or individuals who foresee future punishment from their co-players may attempt to formulate complex and unbounded strategic plans to out-compete their cooperators (Proto et al., 2019), it may be proposed that individuals' propensity to conditionally cooperate is condi-

[5] This type of decision solicitation, the strategy method, has been scrutinised for its dependency on an individual's ability to reveal their 'hypothetical character' (Brandts and Charness, 2011) with particular emphasis on the time inconsistency of individual propensity to punish. Yet, it has become common practice for studies employing this elicitation method to show the validity of their elicitation method under replicated results (Romero and Rosokha, 2018).

[6] An interesting study of such was conducted by Martin et al. (2014) who pitted chimpanzees against humans in a game of matching pennies. The Nash Equilibrium strategy for this game was to randomise perfectly between two decisions. The authors found that while chimps, albeit with experience, honed in on this strategy, humans struggled to implement such an unbiased strategy due to our innate yearn for imitation. See Kahneman (2013) and Henrich (2016) for comprehensive discussions on this.

tional on the context of the interaction. Despite the presence of some research on the conditions under which conditional cooperation arises, empirical evidence of dynamic strategies is not common. Instead, research has mostly focused on the conditions which cause individuals to develop and adjust discrete strategies under two main paradigms as suggested by Bigoni et al. (2016). First, different groups of individuals may respond differently to specific incentives and second, the same individuals may respond to evolving incentives through different strategy choices.

2.2 Existing studies on the instability of cooperation strategies

While personal identifying factors such as gender and nationality have been argued insufficient in their capacity to predict individual cooperative phenotypes (Drouvelis et al., 2021; Rossetti et al., forthcoming), an alternative theory is that effects of these factors on reciprocation may instead be dependent on the environment (Balliet et al., 2011). This theory suggests that individuals may exploit stereotypes or demonstrate instinctive behaviour for material gain; thus gender for instance may instead trigger environmentally dependent instrumental reciprocation. Work by Tognetti and co-authors (2012) has been pivotal on bridging this theory with economic experiments. Their research in rural Senegal identified the sexual motivation of male cooperation, coining the behaviour as competitive cooperation. That is: males cooperate in order to signal their sexual attractiveness when viable female mates are present.

A later study in a similar community confirmed that this effect was driven by single men and was shown to exist only when men are under observation by viable women or village elders who could influence mate allocation (Tognetti et al., 2016). This would suggest that men are more motivated to instrumentally reciprocate in anticipation of future sexual payoff. Kumar et al. (2021) later observed in their prisoner's dilemma experiment that men are more willing to cooperate than women in gender-heterogeneous games and suggested that sexual competition and increased trust in seemingly wonderful women[7] may be a factor here. However, this study concluded by agreeing with the literature consensus: that gender is limited in its ability to identify potential reciprocators on average (see Rossetti et al. (forthcoming) for a review of the literature on cooperation among the genders). Kumar et al. (2021) proposed that this is since gender is prescribed by nature whereas anthropological research implies that cooperative tendencies are culturally adopted (Henrich and Henrich, 2007).

This reasoning would explain why Kumar et al. (2021) found that they could account for a greater difference in reciprocal tendencies using their cultural identifying attribute, in this case nationality, than their non-cultural identifying attribute, gender. Cultural cues are valuable since they imply appropriated individual preferences for reciprocation and individual trustworthiness to repay instrumental reciprocators or punish deviators (Henrich and Henrich, 2007). Political affiliation, for example, has also been shown to imply preferences for cooperation since liberals are believed to be more equality seeking and as such, while conservatives tend to prefer to free ride, liberals tend to cooperate with some non-zero probability and individuals expect them to do so (Balliet et al., 2016; Helénsdotter, 2019). Similarly, religiosity has been shown to be correlated with unconditional cooperation and intrinsic reciprocity (Norenzayan and Shariff, 2008; Billingsley et al., 2018). Further, those who express religious beliefs via symbols are believed as being more pro-social (Mccullough et al., 2016; Potoms and Truyts, 2016). Cultural cues hence may act like signals of intent and thus mirror the observed success of communication in achieving cooperative equilibria (Balliet, 2009). They may also reinstate an instrumental reciprocator's belief that their contemporaneous sacrifice

[7]The term 'wonderful women' comes from the hypothesis studied by Eagly and Mladinic (1994) which proposes that women are more trustworthy and generous than men.

will serve their long-run motivation given that they are interacting with a potential reciprocator.

These arguments suggest that individual identifying factors are not only important for one's propensity to reciprocate but may also affect one's belief of other players' likelihood of contingent cooperation granted interactions are not anonymous. An insightful empirical study to question this hypothesis was performed by Bigoni et al. (2016) who conducted an experiment of repeated public goods games within a community in Northern Italy, and then within a community in Southern Italy. They found that only the northern community coordinated on a cooperative outcome and proposed that this difference was driven by a social capital wedge such that individuals in the south were less culturally attune to cooperative interactions and thus the presence of fellow Southerners triggered a common belief of low cooperation. The notion that individuals interpret incentives within experiments using their real life experiences and stereotypes has long been supported by the literature. This would explain robust results such as economic students free-riding more frequently than students of other disciplines (Marwell and Ames, 1981); or that individuals who are used to interacting with others in their occupation cooperate more (Gneezy et al., 2016). These studies suggested that since conditional cooperation strategies, by definition, are constructed out of belief of future gains, the stronger one believes in the prospect of mutual reciprocation, the more likely one is to employ a conditional cooperative strategy (Ackermann and Murphy, 2019). Research focused on policy has therefore focused on how these beliefs can be moulded to facilitate widespread cooperation.

One strand of the literature which has implicitly studied imposed beliefs has systematically varied game parameters to observe how individual cooperation strategies evolve. As aforementioned, a necessary condition for instrumental reciprocity is that one believes in the potential realisation of future gains from one's current sacrifices, and as such it is probable that one will have the opportunity to interact again (Roth and Murnighan, 1978). The measurable discount factor has accordingly been used in experimental research to quantify the probability weighted value that one places on future interactions and to show that increasing this parameter, increases the rate of cooperation (Murnighan and Roth, 1983; Dal Bó, 2005; Fréchette and Yuksel, 2016). Other studies have instead framed the necessary condition in terms of the probability weighted pecuniary rewards and punishments from the payoff matrix. As such the necessary condition can be framed as ensuring that pecuniary payoffs from mutual cooperation exceed the one-time benefit of defection and subsequent punishments (Mengel, 2018).

Following theories of threshold strategies which propose that individuals cooperate given the satisfaction of this necessary condition, several studies have noted that systematically altering game payoffs alters the level of instrumental reciprocity. For example, Nikiforakis and Normann (2007) proposed that if the payoff to defection is high relative to that of mutual cooperation, then individuals stick to defection strategies and this is strategically expected so cooperation does not take off. Reuben and Suetens (2011) also demonstrated that as mutual cooperation becomes more lucrative in an infinitely repeated prisoner's dilemma, individuals tend to coordinate on instrumental cooperation strategies.

While these results, together with the equilibrium condition, provide important intuition as to why conditional cooperators emerge, it may not always be the case that this necessary condition is satisfied in expectation, or that the individual even recognises the forward-looking incentive to engage in cooperative behaviour. Hence, it is fruitful to identify the static strategies of individuals which are implemented when the future is not sufficiently weighted or payoffs are not in favour of cooperation.

By employing Dal Bó and Fréchette's (2011) method of computing individual strategies, Dreber et al. (2014) and Arechar et al. (2018) found that many individuals do switch from conditional cooperation to unconditional defection as cooperation becomes a non-equilibrium choice. Embrey et al. (2018) supported the above by observing clear threshold strategies employed by subjects as prisoner dilemma payoffs are adjusted to accommodate the critical value necessary for the equilibrium condition. These authors elicit individual propensities to unconditionally cooperate through an initial dictator game to further show that most of the cooperation remaining under non-equilibrium conditions can be accounted for by altruistic tendencies. Individuals who presented the most pro-social inclinations were most likely to cooperate even when the threshold condition above was not satisfied and thus instrumental reciprocity was not an equilibrium condition. Embrey et al. (2018) exploited their use of multiple supergames to propose that individuals may also learn to unconditionally cooperate under non-equilibrium conditions to maintain socially optimal cooperation. Further research into this has shown that less pro-socials individuals in fact may learn to be patient in non-equilibrium conditions and if this is the case, continued cooperation may also be defined as forward looking instrumental reciprocation (Kurokawa et al., 2018; Deb and González-Díaz, 2019).

From the above, it could be proposed that for any cooperation to occur outside of private equilibrium and for individuals to maintain some strategy involving reciprocity, some proportion of the population needs to be sufficiently pro-social or patient and some proportion of the population needs to believe that this is the case. Thus, in contrast to Kandori's (1992) argument that heterogeneity can lead to a breakdown of cooperation, heterogeneity, at least of patient reciprocal motivations, may be fruitful to the maintenance of cooperative norms in a population. These arguments lead to the hypothesis that group dynamics are integral to the evolution of conditionally cooperative strategies.

To discuss group dynamics, one must first identify the mechanism behind group establishment. A common definition stems from early research into social identity by Tajfel et al. (1971) who showed that individuals are prone to social categorisation and behave according to the expectations prescribed to these categories. Akerlof and Kranton (2000) incorporated this conjecture into a utility model such that individuals derive utility from adhering to these behavioural expectations. This is supported by an eloquent description from Bicchieri (1990) of the mechanism under which one can observe group contingent behaviour given that a social norm is established and believed in by group members. Bicchieri (1990) formally defined any deviation from such a norm as a violation of one's intrinsic identity and this has been the basis for many models of cooperation under established social norms (see Brock and Durlauf (2001); Bernheim (1994); Camerer and Ho (1999)). While this theory has until recently maintained some distance from the theory of reciprocity, norm dynamics have been shown to influence the degree to which one may employ strategic and conditional strategies in cooperation dilemmas (Smerdon et al., 2016).

One component of group configuration that may influence conditional cooperation which has attracted significant interest is the size of the group. It is the consensus of the empirical literature that large groups inhibit widespread cooperation, however the reason for this stylised fact is debated. Bicchieri (1990) initially proposed that individuals are proportionally insignificant in large groups such that individual defection can go undetected and perhaps unpunished which renders instrumental and forward looking cooperation fruitless. This is supported by Kok et al. (2020) who observed groups of varying compositions in rural Tanzania to show that smaller groups are able to establish cooperative norms more effectively since they facilitate regular monitoring. However, this

theory is contested by two opposing theoretical hypotheses. First, in large salient groups with strong norms, an individual defector may be more identifiable since they are a unique anomalous observation (Traxler and Spichtig, 2010). Second, strong social norms can be made stable to invasion if individuals adopt some leniency to their cooperative strategies (Dal Bó and Fréchette, 2011). Duffy and Xie (2016) supported this latter rebuttal upon observing that reducing the difference between cooperation and defection payoffs may increase group-wide leniency and thus the stability of cooperative equilibria even in large groups with non-uniformly distributed patience. This underlines the interdependence of interaction parameters in determining the strategic outcome.

The study of heterogeneous groups is topically of interest since modern society presents various melting-pot scenarios. As demonstrated by Dreber et al. (2014), heterogeneity may be socially serving if it aids pro-social cooperators to influence the skeptical reciprocator. Yu et al. (2015) modelled cooperation as utility maximising so long as some pro-social agents are present which forces the average strategy and thus implied social norm to be distinguishable from the polar case of mutual and perpetual defection. This result is supported by multiple earlier studies which did not explicitly model social norms but showed that groups consisting of a few pro-social norm setters can force a coordination on cooperation should they threaten to punish. For instance Fehr and Gächter (2000*a*) saw that punishment in the public goods game inflicted upon those who deviate from the average contribution supports some conformity effect. Moreover, Bernhard et al. (2006) showed that punishment is stronger when a violation affects identifying-group members and thus the credible threat of punishment within well-defined groups can achieve group-wide cooperation if members are accustomed to group-level norms.

Realpe-Gómez et al. (2019) formally defined a model that fully integrates the effect of a social norm of cooperation into the theory of reciprocity by defining an intrinsic trade-off between one's individual material payoff and one's socially derived status according to a norm. This model, termed the Experience Weighted Attraction with Norm Psychology (EWANP) model, is rather novel in its applications but is adapted from the earlier model of Camerer and Ho (1999), who constructed individual utility as a weighted linear sum of material gains, their *individual drive*, and of the degree to which their action is in accordance with some norm, their *normative drive*. This model serves a number of interesting purposes and the basis for this thesis.

First, the model allows a continuous measurement of some social norm and therefore facilitates the modelling of reciprocal behaviour in n-person games with heterogeneous individuals. Thus the model predicts that norms can vary in strength, salience, and by the degree to which groups abide. This permits the empirical support of research that has identified the power of salient social norms in achieving cooperative equilibria. For example, Charness et al. (2007) provided evidence that the salience of group identity increases the strategic dexterity of group members. Charness et al. (2014) demonstrated that a preliminary team building exercise can improve a group's ability to converge on a cooperative strategy but only when the whole team has performed the exercise, thus underlying the need for strong and salient norms. Kurokawa et al. (2018) also studied repeated n-person prisoner dilemma games under a model of dynamic strategies to show that as a greater number of players cooperate, a social norm increases in strength which raises the marginal benefit of cooperation and facilitates the group convergence on a pro-social action. Furthermore, Nagatsu et al. (2018) showed that when groups are transparently homogeneous and a strong social norm is set for cooperation, even natural free-riders can be incentivised to cooperate. These articles demonstrated the value in accounting for non-discrete group composition and social norm effects in the dynamics of strategy implementation.

Second, the model permits an extension first proposed by Bernheim (1994) which varies an individual's sensitivity to the norm. This extension allows a continuous measurement of one's willingness to reciprocate. Unconditional strategies hence have a zero sensitivity to social norms and as such individual utility would be solely realised from one's material payoff. The more sensitive an individual is towards the social norm, the more an individual is likely to reciprocate and as such this extension relaxes the assumption of a discrete individual phenotype of reciprocation. Furthermore, since this sensitivity interacts directly with the social punishment, individuals may display different sensitivities to varying action spaces which permits an evolving or situational unstable phenotype. An early proposition of this method to define degrees of conditional cooperation was presented by Traxler and Spichtig (2010) who focused on the evolution of norm sensitivity in social environments of heterogeneous players. They defined conditional cooperation in this case as the periodic norm conforming strategy which is dominant and evolutionarily stable. The conclusion of their paper is therefore that under continuous parameters, the dominant strategy is to adapt to the specific environment that one expects to be realised.

Critically, this continuous measurement of norm sensitivity facilitates the empirical identification of conditional cooperation, even when widespread defection is observed. Thus, while the Dal Bó and Fréchette (2011) method utilised by Dreber et al. (2014) and Arechar et al. (2018) identifies individual strategies according to discrete behavioural observations, the EWANP model can distinguish between a continuum of conditional cooperators of limited patience responding to always-defectors.

Third, unlike models of stochastic strategies such as that of Breitmoser (2015), this model can predict a complete pure dynamic equilibrium strategy for individuals upon observing a social environment and can therefore show the evolution of individual strategy according to social norm sensitivity and individual private incentives.

This model, using group dynamics and a holistic definition of individual incentives, is an appropriate advance in the avenue of research into cooperation. However, this theory has not yet been applied empirically. Hence, to the best of my knowledge, this is the first empirical study to employ this type of model to study pure strategy dynamics with variable social norm sensitivity and evolving environmental parameters.

3 Theoretical model

The functional form of this thesis' theoretical model is derived from the Exchange Weighted Attractiveness with Norm Psychology model first proposed by Camerer and Ho (1999) and finalised by Realpe-Gómez et al. (2019). This model is chosen in order to analyse the linear separation of one's private utility derived from private payoffs of individual actions and one's social utility derived from one's adherence to a known social norm. The specification of the social cost function and the derivation of equilibrium conditions and drawn from the discrete choice model of Brock and Durlauf (2001) and the generalisations outlined by Smerdon et al. (2016). I have additionally modified these definitions to include a specific definition of the private payoff function.

Consider a neighbourhood of N individuals who are part of the wider population $\mathcal{N}$ such that $N \subset \mathcal{N}$. Each individual in each neighbourhood is faced with a binary choice: one can either **cooperate**, which requires the payment of a cost c, or one can **defect** and free-ride. Denote the

variable ω_i as the choice variable where $\omega_i = 1$ when the individual cooperates and $\omega_i = -1$ when the individual defects. Further ω_{-i} represents the vector of actions that i believes other players will take.

Let the individual's utility function be the following:

$$u(\omega_i, \omega_{-i}) = v(\omega_i) + \lambda_i S(\omega_i, \omega_{-i}) + \epsilon_i(\omega_i) - (\frac{\omega_i + 1}{2})c \tag{1}$$

Where $v(\omega_i)$ represents the private common value derived by each individual who chooses some behaviour. Assume for simplicity that $v(1) \geq 0$ and $v(-1) = 0$ such that the private common benefit of cooperating is always at least as good as the private common benefit of defecting. c is the cost of cooperating. $S(\cdot)$ represents the social cost function which punishes the individual when they deviate from the average choice of other individuals, $\bar{m}_i^e = \frac{1}{N-1}\sum_{j \neq i} \omega_j$. This component introduces a strategic element to the payoff function which can be easily generalised for an arbitrarily large number of players. λ_i represents the weight that individual i puts on the social cost function in their utility derivation. Thus a higher λ_i represents a more sensitive individual with respect to their social standing. This parameter crucially allows the observation of a continuous degree of conditional cooperation. Let $\epsilon_i(\omega_i)$ be the stochastic component of the utility which is dependent on player and choice. It is assumed that for each player $\epsilon_i(1) - \epsilon_i(-1)$ has a known distribution, $F(\cdot)$, with mean of zero.

Assume the functional form of $v(1)$ to be the following:

$$v(1) = \frac{b}{r+1}$$

Such that r is the risk that one faces when they do not cooperate and the b is the potential maximum private marginal benefit that one achieves from deciding to cooperate.[8] This private incentive is hence an applied benefit-cost ratio function. In order to apply the model to the literature, it is important to assume that $r > 0$ such that the function above is defined and weakly positive. In this study, I refer to $v(1)$ as the private incentive to cooperate and $\frac{1}{v(1)}$ as the external threat.

Following Brock and Durlauf (2001), I assume the form of the social cost function as the following:

$$s(\omega_i, \omega_{-i}) = -\frac{J}{2}(\omega_i - \bar{m}_i^e)^2$$

This form of the social cost function captures a strong conformity effect which punishes individuals at a proportionally larger rate the further their behaviour is from the average choice of other

[8] Recall the recycling example from the introduction. In this case r can be characterised as the environmental cost one feels if they do not recycle and thus neighbourhood-wide cooperation breaks down. b in turn can be defined as the environment benefit one will individually enjoy should they cooperate and help establish a neighbourhood cooperation norm.

individuals. The social parameter, J is the weight that the population as a whole places on social conformity. Thus a higher J indicates a harsher social punishment on a deviator. This social value function hence is equal to zero only when individual i conforms to the expected average choice of others. For simplicity, I assume that J is strictly positive, known and fixed. In the context of indefinite cooperation decisions, J can represent the future punishments that i will receive if they deviate from the norm in the current period.

To derive the equilibrium conditions, characterise the parameter ρ^* as the equilibrium proportion of the population choosing to defect:

$$\rho^* = \frac{1 - m_i^*}{2}$$

Where m_i^* is the average choice of the entire population of i's neighbourhood, $m_i^* = \frac{1}{N}\sum_{i=1}^{m} \omega_i \in [-1, 1]$.

Define $d_i = \epsilon_i(-1) - \epsilon_i(1) - v(1) + c$ as the difference between the private value of an defecting player and the private value of the cooperating player. d_i is hence $i's$ net benefit from defecting in absence of social effects. Thus whether or not defecting is a private equilibrium choice depends on whether d_i is at least as good as some threshold value denoted Q^*. Following the specification of the utility function, Q^* is dependent on the individual's beliefs on other's behaviour, the social parameter J and the individual's social weighting λ_i. Accordingly, this condition can be rephrased as the classic condition for cooperation: cooperating is a private equilibrium if the private payoff from defecting now d_i is less than the loss in future payoffs due to social punishment, Q^*.

Let p be the probability that any individual draws $d_i < Q^*$ and thus chooses to cooperate. Solving for this condition such that $U(1) > U(-1)$, I can define $Q^* = 2\lambda_i J\bar{m}_i^e$ and therefore I can define p as:

$$p = P(\epsilon_i(-1) - \epsilon_i(1) < 2\lambda_i J\bar{m}_i^e - c + v(1)) = F(2\lambda_i J\bar{m}_i^e - c + v(1))$$

In equilibrium, it must be that that $\omega_i = \bar{m}_i^e = \bar{m}_i^*$ such that the average choice of all players including any individual i is equal to the expected average choice of all players except for i. This results in an endogenous solution such that:

$$p^* = F(2\lambda J\bar{m}_i^e - c + v(1)) \tag{2}$$

Where λ is now the neighbourhood average social cost weighting factor and measures the degree of conditional cooperation the average individual exhibits. Additionally, it is possible to solve for $\bar{m}_i^*$ to be the following. I relegate the proof of this to Appendix 1.

$$\bar{m}_i^* = 2F(2\lambda J\bar{m}_i^* - c + v(1)) - 1 \tag{3}$$

According to Brock and Durlauf (2001), at least one equilibria must exist to the above and multiple can exist when λJ is relatively large with respect to to the common value, $v(1)$. This is the case since individuals with non-zero values of λ exhibit some degree of conditional cooperation and so their tendency to cooperate or defect is dependent on the social norm. Those with zero values of λ choose their strategies unconditional of social norms and so only one equilibrium can be realised. In this case, it is possible to characterise two stable equilibria around each decision node i.e. $\rho_D^* \approx 1$ and $\rho_C^* \approx 0$. Under this model, the equilibrium defined by ρ_C^*, where almost all agents cooperate, is the total welfare maximising equilibrium if $v(1) - c > 0$ (Brock and Durlauf, 2001). In other words, ignoring all social effects, it is in each individual's best interest to cooperate if the net private benefit of cooperating is positive. When under a social norm of cooperation, this equilibrium will also be a social welfare maximising equilibrium since any socially sensitive individual's unilateral deviation will cause each individual to feel a social cost since the average choice is less than 1.

In the case where $v(1) - c < 0$, then the net private benefit of defecting is positive and it is in each individual's best interest to defect from cooperation. Thus the lower the private common value of cooperating, the less likely there to be a cooperating equilibrium. The inclusion of social effects allows the welfare maximising equilibrium to be defined as ρ_C^* even when $v(1) - c < 0$ as long as a social norm for cooperation is present and individuals are sensitive to their social status however, the relative potential for a social norm to influence cooperation diminishes as $v(1) - c$ tends to negative infinity. Thus when the private equilibrium choice is to defect, social welfare may still be maximised if conditional cooperation occurs under a cooperative norm.

3.1 Hypotheses

The model clearly defines that the individual utility is increasing in one's private incentive, $v(1)$ and in one's accordance to the social norm. Since private and social incentives are additively related to utility, I propose that for any level of social incentives, the marginal utility of cooperating increases as the private benefit of cooperating increases. Further, this suggests that at any private benefit of cooperation, the marginal utility from cooperating increases if one cooperates and others are also cooperating. This intuition results in the first hypothesis.

H1a: Individual marginal utility from cooperating increases as the private common benefit increases.

H1b: Individual marginal utility from cooperating increases as the number of other individuals in i's neighbourhood cooperate.

Hypothesis one (a) suggests that as the private common benefit to cooperation decreases, one has a lower marginal per capita return to their decision to cooperate given some level of social incentive. Hypothesis one (b) proposes that the marginal return to cooperation increases as the social incentives increase. A further interpretation of hypothesis one (a) is that the attractiveness of cooperation and thus one's incentive to cooperate decreases as the level of the private incentive to do so increases. This is formally defined by hypothesis two.

H2: The incentive to cooperate increases as the private common benefit increases.

Despite this, my model predicts a threshold $v(1) - c = 0$ where the two polar equilibria $\{\rho^A, \rho^C\}$

are equally attractive absent of social incentives. However social incentives may continue to play a positive role in an individual's incentive to cooperate even when cooperation is not a private equilibrium. This is the result of Dreber et al. (2014), which my model also predicts. First, the model predicts that with private incentives below the threshold, social incentives can increase one's propensity to cooperate. Second, the model predicts that as the private common benefit from cooperation decreases, the social incentive to cooperate also diminishes in its capability of persuading a cooperative norm since the individual may expect fewer individuals to cooperate which further lowers m_i^*. This sequential mechanism is summarised in hypothesis three.

H3: Individuals are more likely to cooperate when social incentives increase however the marginal effect of social incentives decreases as private incentives fall.

Hypothesis three can be interpreted as saying that individuals are more likely to conditionally cooperate the greater the proportion of cooperators are present. However as the private incentive to cooperate diminishes, the propensity to conditionally cooperate diminishes.

Finally, the model predicts that groups who are on average more sensitive to the social norms will achieve higher cooperation rates when the social norm for cooperation is present due to a higher value of λ. I propose that under the concept of in-group favouritism (Tajfel et al., 1971), it may be the case that groups of individuals who are close to the average group identity will be more sensitive to their accordance with the average choice of their group. This suggests that the closer the group feels to each other, the closer they will be to perfect contingent cooperators. This is formulated in hypothesis four.

H4: Individuals are more likely to cooperate when their identifying attributes are closer to their group's average identity.

4 Data

4.1 Background

The data used is contestant decisions and characteristics from the 40 seasons of the US game show series *Survivor*. During this series, between sixteen and twenty contestants, who are personally unknown to each other, are introduced to a remote location and split into multiple tribes. These tribes compete against each other in intermediate public goods games (IPGG)[9] for luxuries while working together in their respective groups to find food and shelter. In each round, a competition is performed between the two tribes where the losing tribe must vote among themselves to eliminate one of their own members. The player receiving the modal votes leaves the competition.

During the game, tribes can swap members at the producers' discretion. At around the half way

[9]An intermediate public goods game is defined as a classic public goods game that occurs mid-way through a round. Individuals must invest some effort into the pool of group effort in order to compete against a rival tribe for a reward. Each individual has an incentive to defect since they would still benefit from the effort of their fellow tribe members without the effort cost. Since individual actions may be visible though, it may be privately optimal to invest maximum effort. The purpose of the IPGG in this study is to measure group trust and morale as expanded on in the next section.

point, a grand merger takes place where all remaining contestants join to form one single tribe. The game then continues where at the end of each round, all players cast a vote for who they wish to eliminate.

At any point throughout the series, contestants have the opportunity to form an alliance. An alliance is a set of rules among the cooperating players dictating that they should not vote for each other and that they should all vote for a specified target should they be faced with the decision of whom to vote for. Contestants are not however, primed to form these alliances and do so at their own free will.

The players who are eliminated then regroup at the end of the game to vote for a Sole Survivor among the final remaining two or three contestants. The Sole Survivor then receives a cash prize of one million dollars. Runners up claim a prize which decreases in the position at which they finish.

The dataset is composed of three main databases. The first database is a viewer-collected and peer reviewed database accessible at survivor.fandom.com. This database contains contestant characteristics such as age, sex, and occupation, as well as information on voting histories, alliance formation and challenge success rates. The second database supplements the above with contestant specific information from a viewer-collected database. This database is publicly available at [https://CRAN.R-project.org/package=survivoR]. The third database consists of viewer-transcribed contestant confessionals which allow contestants to share their thought processes with producers and viewers. Note that these confessionals are not observable by other contestants. This database is used to complement the above with player intentions on cooperation, however, it is rarely used in addition to the above since many alliance descriptions in the primary database describe this[10]. Each season of the series has between 16 and 20 players and I collect data from three periods before and after the mid-game merger event. This panel length was chosen primarily for time reasons since coding the decisions of contestants is cumbersome. Further the merger event provides the majority of variation in the independent variables and so is the most important time period to consider. Thus, at a baseline, I have potential for up to 2160 individual observations from 400 individuals.

Additionally, series' synopses are gathered using Wikipedia in order to obtain information relating to the exclusion criteria.

4.2 Application to the model

To apply the data to my model, I should begin with the exact definition of an interaction and a neighbourhood. A player interacts in time period t if they have the opportunity to cast a vote in that period. They may not have this opportunity if the player has been eliminated; if the period is prior to the merger and they win an 'immunity challenge', or if a player is medically evacuated and the vote is called off.

A neighbourhood is a selection of players who are playing together in the same tribe who are not aligned against one another and who share a common threat. Explicitly, a player's neighbourhood ID is derived via the following criteria: individuals are assigned to one alliance id, which is either the alliance that they join at one point during the game or the alliance that is composed of players which they had the most interaction with prior to the merge. The neighbourhood is then the group of individuals who are assigned to the same alliance id who interact with each other at

[10]This database was accessed at [https://drive.google.com/drive/folders/0B8Xzl82K1TP8fmItS2RoYWUxeW1YSmZoUXVQSldNMTJnUEVSV1Zvd2xYaFpLYnViOWJ1RXM?resourcekey=0-lnqLgepahAhBF8fOjYeVrw&usp=sharing]

time t. This means that at the merge, the neighbourhood may grow to include previously separated tribal members following a tribe reshuffle.

In each period, each player faces a utility maximisation problem which is to maximise their probability weighted payoff from the game. This payoff is measured by one's expected position of elimination in the game which is endogenously determined by the behaviours of other players who may act cooperatively with the individual or may pose a threat to the individual neighbourhood.

Throughout this paper, I will define cooperation as the situation where one is an explicit member of an alliance and votes in accordance with the group level agreed rule. This rule is normally unambiguous but in certain occasions such as during split-voting strategies, inspection of transcripts and synopses is necessary. A cooperative action is presumed to act in favour of the neighbourhood itself. This act may also be an individual utility-maximising action also.

The parameters J and λ_i both relate to intra-neighbourhood interactions and together represent the individually weighted social cost of deviating from the neighbourhood norm. A player may be rather vulnerable to being eliminated if the neighbourhood cooperates and the player in question does not, since the neighbourhood may then decide to act in revenge against non-cooperating members. However, under extreme preferences or with multiple rival neighbourhoods present, a single non-cooperator may be less socially vulnerable to this utility loss.

A neighbourhood faces a common threat if there exists a rival alliance in an opposing neighbourhood. Since each individual can cast one vote and the player with the modal vote is eliminated then the relative threat of a rival alliance is the number of members in this rival alliance divided by the number of individuals in individual i's own neighbourhood. As such this threat is weighted by its manageability. A high relative threat does not necessarily mean that the rival alliance is absolutely large but that the return to cooperation is very low. A value of one hence suggests that the number of opposing players equals the potential size of one's own alliance. It is worthy to note that this measure does not incorporate whether rival players specifically cooperate since this is only visible after one has made a decision to cooperate. Instead, it is assumed to be far more visible whether a rival player belongs to a specific alliance than if one abides by its rules in voting due to previous decisions and to assumed limited communication between rival players.

The utilisation of TV game show data, while not novel, remains rather uncommon in economic literature. This is likely due to a number of reasons such as lack of complete observation, complex interactions and individual-time heterogeneity. Most importantly however, since the use of secondary data for behavioural experiments does not come with the high standard of control realised in the laboratory, one is not able to draw concrete causal conclusions from the experiment. Nonetheless, there are numerous advantages to using this data. First, the payoffs to players are far greater than in standard laboratory experiments and so by exiting early, an individual incurs a very significant loss of potential earnings. Thus with these payoffs, it is more probable that behaviour is more relevant to real life, high-risk scenarios where only rational non-altruistic strategies are most likely to be revealed (Post et al., 2008; Rabin, 1993). Nonetheless, individual strategic behaviour in the game show has been shown to follow economically predictable patterns (Karlan, 2017). Second, individuals repeatedly interact with one another in settings which mimic survival games where payoff streams are contingent on strategic behaviour. This is a setting that is similar to one in which anthropologists argue cooperation is natural for humans (Bowles and Gintis, 2011). As such, individuals are less likely to suffer from abstract bias and react according to individual instincts. Third,

I am able to identify specific neighbourhoods which interact using the game rules. Finally, since the series is long-running and international, the potential size of the data set is rather large and growing.

Data was collected on individual decisions three periods before and three periods after the mid-game merger regardless of if the player interacted in each periods. However, the rules of the game are altered in some special seasons. This renders some repetitions of the game not suitable for analysis. Certain criteria were ex-ante defined which excluded data from specific seasons. In particular the following conditions must be met on the season level for a season to have been included in the panel data set:

1. There must be at least 2 separate tribes who merge and each tribe must have a minimum of 2 members at merge. This ensures sufficient variation in individual data.

2. Contestants cannot be known to each other prior to entering the competition. If this does not hold, this may alter the willingness of participants to cooperate and creates space for interaction outside of the TV show, errors in the measurement of variables, and therefore bias in the coefficient estimates.

Following the implementation of the two above criterion, there remains observations from 384 individuals over an average of 3.28 periods. These criteria suit the implementation of my first model which models the drop-out rate of contestants according to individual incentives. However controlling for drop-out rates (attrition) and instability in alliances in further models is necessary to limit the prospect of attrition bias (Little and Rubin, 1987). Thus, I specified two further criteria to be satisfied. Each criterion and the reasons for their implementation are described below.

1. Quasi-balance criterion: There are three reasons for which an individual does not interact and thus has no observation in a period. First, an individual could have been eliminated and is no longer in the game. Second, prior to the merger, it is possible that the individual is part of a tribe which won an immunity challenge and thus they do not need to vote. Third, a round of interactions is called off due to a medical evacuation. When using panel data, non-random missing values (attrition) is problematic since it may trigger attrition bias in the estimated coefficients (Little and Rubin, 1987). A missing value before the merger is somewhat stochastic for an individual that survives the merger since this indicates that their tribe did not have to vote, and so the modelling procedure can tolerate this. Missing values post-merger however normally signal that the individual has been eliminated since there is no possibility to escape the vote. Elimination is not a random event in face of the decision to cooperate since according to the theoretical model, cooperation may enhance one's probability of advancement. Sometimes, however, a round does not result in an observation due to medical evacuations and in this case, the individual is still alive in the game. I treat this event as random. Hence, I ensure that individuals have at least one interaction prior to the merger and interact in the third period after the merger: they survive my periods of observation and any non-observation is random. By excluding attrition from this subsample, I must however, be precise with my interpretation: here I only observe individuals who survive through periods of observation.

   This criterion is not employed for model one described in the next section since this model relies on attrition of subjects for variation in the dependent variable: the probability of continuing to the next period. Specifically, in the first model, I require that some individuals do not continue to the next period in order to model why this is the case. However, for all subsequent models, the balancing exclusion criteria is necessary in order to ensure comparability between

individuals as they choose to cooperate over different time periods. If there is a positive relationship between cooperating and game longevity and the model was not somewhat balanced, I would expect a strong bias towards cooperation at the end of the time frame regardless. This is because the missing data will not be random according to the dependent variable and as such, I would expect an inflated error term.

2. Stability criterion: Individuals cannot flip alliances or flip between cooperative relationships. If I allow for individuals to join any alliance then the choice set at each period becomes exponentially large and the model becomes extremely vast and complex. Importantly, this criterion uncouples any serial relationship between the private and social incentives caused by individuals moving between alliances and allows for the assumption of an exogenous external threat. If individuals are allowed to flip alliances then they could in effect choose the threat that is best for them and the private threat becomes endogenously malleable which would render my results biased.

5 Empirical methods

The empirical methods, unless stated otherwise, were defined before performing regression analyses. However, the exact method was devised during the data collection phase since the exact attributes of the data were unknown prior to this. Before discussing the specifics of the modelling procedures, it is worth discussing two points: data ethics and identification.

5.1 Data ethics

Under GDPR regulations, it is not permitted to collect, manipulate, and process personal data[11] (EU, 2016). Clearly, by the definition of this data set, all individuals are identifiable since the data set is publicly retrievable at any point. To combat this, I employ a methodological procedure to anonymise the decision data set under a specific key. This same key is used to anonymise individuals and their sensitive data in the transcript data set. Instead of directly recording sensitive data e.g. sex, I will code a different variable which computes the likeness of individuals in the neighbourhood N_t. For example, to compute the sex similarity among a group, a variable $sex\_concentration_i$ which is the relative number of individuals in i's neighbourhood who are the same sex. A similar method is used to compute the relation of individual i's age to the group average, along with social economic status and hometown.

Further, GDPR documents state that *'any information that is manifestly made public is acceptable to process including sensitive data'* (EU, 2016). The topic of debate is hence whether data made public via participation in a TV show is **manifestly** made public. The answer is not unambiguous. That contestants do enter the competition knowing and agreeing to that every decision and personal detail may be televised to millions of viewers does provide pretence to argue that they purposefully agree to make their data public. However, the final issue is whether this manifestation can be interpreted as permission to analyse this data. Under an internal ethics microscope, this is not problematic since my research is for a thesis and not a research paper. From a data protection standpoint, it is difficult to argue for a specific result without a scrutinous analysis of contestant contracts. For this reason, I cautiously continue to use anonymisation techniques.

[11] Data on individuals that allow them to be identified

5.2 Identification

The identification method employed uses a two-way fixed effect linear probability model. This model has been chosen due to the assumption that there will be some covariance between both individual effects and period of game effects and the independent variables, thus it is appropriate to include individual and time effects in the intercept term. The correlation between individual effects and independent variables is rather unambiguous since some individuals have a higher baseline probability of conditionally cooperating or succeeding and this will correlate with average sensitivity to social settings and individual propensity to conditionally cooperate. The correlation between the period of the game effects and the independent variables is also not a radical assumption. Time in the game poses three sources of covariance with the independent variables. First, the longer the individual has been in the game, the longer they have been able to interact with the game and fellow participants and the more likely that individuals may learn that cooperation is feasible and beneficial. The literature widely supports that experience in games enables individuals to learn how to cooperate through improved beliefs of other contestants' propensity to reciprocate (Dal Bó and Fréchette, 2011; Balliet et al., 2016; Embrey et al., 2018) or through accustoming themselves with the game procedure (Mengel, 2018). Second, the longer an individual has been in the game, the more likely the merger has occurred which presents a shock to the independent variables. Finally, the longer an individual has been in the game, the fewer players there will be in the entire game creating a more competitive landscape and perhaps facilitating neighbourhood cooperation (Bicchieri, 1990; Kok et al., 2020). The third factor is directly controlled for in the independent variables. Further, I control for the second point using a dummy variable which equals one if the merger has happened. The first source is controlled for by including period of game fixed effects.

A linear probability model is chosen over a logit model in order to allow for the inclusion of interaction variables and to facilitate effect interpretation. This is because the logit model employs a non-linear transformation so included explanatory variables and their associated coefficients contribute to the estimated probability of an event in a multiplicative manner. Therefore, the inclusion of an interaction effect has a scale-dependent impact on the logit model (Hosmer et al., 2013). While the linear probability model has the key limitation that predictions may lie outside of the $[0, 1]$ interval, the use of this model is accepted in the literature in studying effects.

The independent variables of interest are time specific variables and time varying controls. However, given that the data is not derived from random trials or first-hand observations, multiple aspects of endogeneity threaten the analysis. These aspects are limited to three separate issues.

The first issue regards errors in variables which, in this matter, relates to individual decisions being made in consideration of outside consequences beyond the direct monetary payoff to the game. This concern, is partially remedied by the exclusion criteria of not allowing pre-existing relationships to enter the game nor individuals to choose their alliance and thus external threat. Since there is no mechanism to fully control for variable measurement error without violating some form of anonymisation protocol, these observations must be dropped. A discussion on the impact of growing relationships is included below in issue three. Hence, an initial identifying assumption is that individuals in my restricted sample make voting and cooperative decisions based only on the payoff to the game.

The second issue relates to the unobservable (or inability to ethically observe) time-invariant heterogeneity across individuals and individual-invariant time heterogeneity across periods. Most of the present individual heterogeneity relates to time-invariant sources such as race, religion or sexual

orientation whereas specific events in the game such as the merger threaten time heterogeneity. In a laboratory setting, individual factors would not be of concern due to random selection and anonymisation however in the real world, these attributes may be salient and impact decisions and so these must be controlled for in order to maximise the external validity of the experiment design. This is not only difficult to record from transcripts, but is also randomly revealed and immoral to explicitly process according to GDPR. Because of this inaccessible information, the use of cross-sectional analysis or individual pooled analysis is ultimately made inefficient and biased. A solution is to use two-way fixed effect methods which adopt within-individual analysis: using individuals as their own control at different time points and then averaging this over time while controlling for specific groups of periods such as post-merger periods (Hosmer et al., 2013).

There are, however, several limitations to the implementation of fixed effects. The two limitations most worthy of mentioning are the inability to measure time-invariant attributes and the tendency to misinterpret coefficients in absence of consideration for the study design. (Hill et al., 2020). Certain time-invariant identity-related characteristics such as socio-economic status and sex may be important to the result since salient individual attributes may impact cooperative tendencies through in-group bias.[12] To account for this, a vector of control variables Z_{it}^{c} is coded which account for the likeness of the individual i to others within i's neighbourhood at any given time. There are two key benefits of using this likeness control vector. First, it allows the quantitative measurement of the in-group favouritism an individual may feel to those in their tribe at any one time thus controlling for a specific covariate of individual social-norm sensitivity. Second, the vector elements can be included in fixed effect specifications since group composition varies over time.

The final issue of endogeneity relates to excluded control variables which are time or individual variant relating to individual inclination to cooperate. While time- and individual-invariant heterogeneity is dealt with above, evolving preferences are likely to be present and are difficult to control for. In this model, I am, on the one hand, restricted by the data that I can include due to the lack of quantitative scales which classify relationships from transcript data. On the other hand, parsimony requires that I cannot include all relationship parameters since this may result in an overfitted model. Further, the demand for data would extend beyond the scope of the public data set. Viewers of the series may assert that the identification of relationships is important as it may introduce an additional motivation other than the monetary payoff to the individual. To this, I admit the shortcomings of my data set but argue that because the evolution of relationships is not necessarily averaged in one direction, I can formulate an identifying assumption that the aggregate effect of relationships is negligible in the study. I also attempt to control for weak group-inherent preferences towards other players by including a voting history composite control detailed below.

One cause of hesitation in the inclusion of period fixed effects over a more complex time-payoff fixed effect composite is that the magnitude of payoff incentives do change as time progresses. Nonetheless, I assume that there is no discontinuous jump in the proportional increase of payoff by remaining in the contest for one extra period. Hence, I do not expect this to impact the external validity of the results through a newly introduced channel of endogeneity.

[12] This effect of interacting pro-socially only with one's in-group is known as parochial altruism (Bowles and Gintis, 2011).

5.3 Control variables

As per Rossetti et al. (forthcoming), I divide the control variables into three sub groups of covariates. The first looks at behavioural cues or rather the past actions of individuals that are observable. The second moves on to personal attributes that are visible such as age or sex. Finally, I include situational cues which refer to the environment under which individuals interact.

The first set of control variables to be discussed relates to historical decisions of players in neighbourhood n, namely voting history and effort history in the intermediate public goods game. This set of control variables therefore aims to control for evolving intra-neighbourhood relationships and group morale. Since both of these attributes may be time invariant or at least contain some inertia, I limit the risk of autocorrelation caused by introducing multiple lagged variables by constructing contemporaneous control variables.

The first control in this vector is an individual-time specific measure for the average relative number of times an individual within the neighbourhood has voted for individual i in the last four periods. The second control exploits the use of intermediate public goods games (IPGG) between voting rounds. In the setting of the series, this IPGG is called the reward challenge. An issue discovered after data collection commenced is that not every period contains a reward challenge and later on in the game, these challenges only credit a sole winner; so the control variable may not serve its purpose. For this reason, the variable was not registered when a winning player was unique. Further, to include this measure when it can only be recorded in certain periods requires the assumption that the loss of a reward challenge has statistically the same effect as there being no reward challenge to play. Materially in the game, the outcomes are identical. However psychologically, a player may feel an inherent utility decrease after a challenge loss and accordingly the neighbourhood may experience a loss in morale. Therefore, I also include a composite variable which indicates the maximum number of rewards challenges a player could have won in the last four periods. The residual of this variable from the variable of indicating the number of reward challenges won by the individual in the last four rounds, is included as a variable proxying for group morale.

I propose that the contemporaneous control composites must account for inertia however. As an example, suppose that in the first round, individuals a and c voted for individual b when all others voted for individual c. While c exited the game, a remained and is still in the neighbourhood along with b. Since the first round however, a has only ever voted for the modal choice and has become close with b. No other remaining player has voted for anyone but the modal in other rounds. The model would predict that since a voted for b in the first round only, that b would be less likely to cooperate than others in b's neighbourhood. Hence, I propose that these controls for voting history and morale should be weighted to assign a higher value to the most recent vote/game and tend to zero after a certain number of periods. I set this number of periods arbitrarily as four, and assign a linear weight which is decreasing from a weight of four for the most recent time period down to a weight of 1 for the fourth recent time period. The same weighting procedure is also assigned to IPGG success and IPGG potential success in order to compute the residual. These measures are then divided by ten such that the maximum possible values for both are one. A value of one for the voting history control would indicate that all individuals in i's neighbourhood have voted for i in each of the past 4 periods. A value of one for the residual IPGG success control would indicate that the individual has lost all of the played IPGGs in the past four periods that i had the opportunity to play with their neighbourhood.

The second set of control variables consists of measures of the individual's similarity to their neighbourhood at some period. The intuition is that as the concentration of specific identities within an interacting group increases, the probability of in-group bias working as a persuasion mechanism for cooperation increases (Tajfel et al., 1971). Specifically, my model predicts that this mechanism is enforced through two mechanisms. First, an increase in group similarity increases one's sensitivity to the cost of deviating from the social norm since one may identify with a group decision (Charness and Chen, 2020). Thus any believed chance that someone will want to cooperate will increase the probability that i will cooperate. Second, more alike individuals will be more likely to have similar preferences and so one can infer the group behavioural norm with greater precision (Nagatsu et al., 2018). This vector does not control for time-invariant differences in baseline incentives to cooperate since this is controlled for by the inclusion of fixed effects.

Some literature has found that cooperation can be facilitated with increased diversity. One example is Parrotta et al. (2014) who find that diversity within certain business teams leads to an increased ability to solve creative problems. A more specific case is highlighted by Tognetti et al. (2016), who find that sexual competition promotes cooperation in mixed groups where men are forced to compete against each other to signal to the women of the group that they are the most viable mate choice. The likeness variables included are the proportion of the neighbourhood who are the same sex as individual i; the proportion who have a higher and a lower socioeconomic class as player i, where socioeconomic class is defined by mapping occupations to the five-class version of Erikson and Goldthorpe's (1992) table; the percentage deviation of player i's age from the neighbourhood average age; and the proportion of the neighbourhood who come from the same geographical division and region as player i. I use nine geographical divisions and four geographical regions according to the United States Census Bureau (2021). Since neighbourhoods are time-invariant in their structure, these control composites permit the control of characteristics that may have otherwise been omitted from a fixed effects analysis.

The third and final set of control variables represent social situation specific measures that are defined in the model. Some seasons include a greater initial number of participants or include double eliminations, thus period fixed effects may not fully control for the total number of players in the game at any one time so it is appropriate to include this as a control variable since more players increases the probability of threat being realised but could decrease the baseline probability that individual i will be eliminated. Finally, a control variable equalling one in periods after the merger and zero before controls for the fact that multiple neighbourhoods will operate together after the merger and as such strategies may change to accommodate for evolving social climates.

5.4 Regression specifications

Through the following models, a panel data analysis is performed observing individual interactions in sequential periods. To account for individual heterogeneity, fixed effects are utilised. However, the issue of individually correlated and heteroskedastic standard errors remains. Cameron and Miller (2015) note that while fixed effects may control for some of the within-individual correlation, they are not a perfect remedy leading to inconsistent coefficient estimations. This merits a specific experiment design which utilises cluster-robust standard errors on the individual level and the use of within-individual computation rather than least squares dummy variables in STATA. Further, any serial correlation that may be present will be accounted for by the use of cluster-robust standard errors provided the number of periods is not too large which is the case in my experiments (Wooldridge, 2013). Additionally, post estimation tests of fixed effects versus random effects require the use

of a cluster-robust modified version of the Hausman test. This is performed by the user written program *xtoverid* which reports a Sargan-Hansen statistic according to a Chi-squared distribution (Schaffer and Stillman, 2010). Finally, it is worth noting that I a priori stipulate the requirement of a significance level of 0.5% to reject a hypothesis and a significance level of 5% will signal a suggested rejection requiring further research as per Benjamin et al. (2018).

5.4.1 Model one

Model one is used to test the first hypothesis of this thesis. First, I construct a baseline model which accounts for private incentives, social incentives and the individual decision to cooperate in the probability that an individual survives a round. The dependent variable is therefore my proxy for individual utility in each round. Second, I apply a modification to account for the interaction between the individual decision to cooperate and private and social incentives separately to test whether an increase in these incentives increases the realised marginal per capita return to cooperation.

I construct a within-individual linear probability model with the dependent variable P_{it} as a binary variable equalling one if that individual i is active in round $t+1$ and zero otherwise. This is my proxy for individual utility since it increases the expected payoff of the game.

The first independent variable of interest here is C_{it}, a binary variable indicating whether individual i cooperates in period t according to my definition of cooperating. To determine the private incentives, I construct a variable using two measures. If n_{it} is the population of i's neighbourhood at time t (the potential maximum marginal benefit of joining a cooperative equilibrium) and R_{it} is the number of individuals not in i's neighbourhood who are in an alliance (the maximum anticipated external risk to the individual if they do not cooperate), then the threat level is considered as:

$$T_{it} = \frac{R_{it}}{n_{it}}$$

Hence, a high relative threat is considered the case where $R > n$ and as such T will be strictly greater than one. Any threat above this threshold indicates the opposition is more numerous than the pool of potential collaborators i has access to and thus the threat is less manageable. I note that this figure can always be defined for an active individual i since $n_{it} \geq 1$ if i is active. Specifically, the parameter T_{it} is the inverse of the private benefit, $v(1)$, as defined in section 3 and applied to the data.

Since the variable for relative threat is constructed using a non-linear function of two covariates, I propose to normalise the distribution of the variable with a logarithmic transformation. This transformation will also aid in the interpretation of the coefficient and ensure that the coefficient captures a linear effect. However, one common drawback of this method is that measurements of zero for the variable cannot be represented after a logarithmic transformation. This requires an additional adjustment to the transformation which does not complement the intuitive advantage of a simple logarithmic transformation. Nonetheless, I continue to construct the new variable $LT_{it} = ln(T_{it}+1)$ to ensures that I am able to represent zero threat values and observe a linear interaction between the components of threat and the dependent variable.

The final independent variable of interest represents the proxy for social incentives, O_{it}. This is

specifically coded as the proportion of individuals in i's neighbourhood who cooperate when i is excluded. I measure social norm effects through contemporaneous cooperation decisions for two reasons. First, unlike in standard laboratory games, individuals are free to communicate and strategise allowing for optimal signalling of cooperative tendencies through communication, which may be more accurate than signalling through one's previous actions (Realpe-Gómez et al., 2019). Second, individuals may not have been able to interact or observe if neighbourhood members in previous rounds did cooperate due to tribe partitions and imperfect visibility of voting. Hence, I assume that individuals within a neighbourhood can make rather precise predictions on their fellow neighbourhood members' tendency to cooperate. I relax this assumption in a robustness check section to include a more observable action.

To address the concern of collinearity between LT_{it} and O_{it} derived from a common denominator component, I proceed in the experiments while testing for this feature using the VIF test and correlation analyses. The results of these tests are reported in the next section.

The below regression characterises the baseline regression used where X_i is the individual fixed effects, Z_{it} is the vector of control variables, and ϵ_{it} is the zero-meaned individual-period error term. In this initial experiment, it does not follow from my intuition to include period fixed effects since I assume that this would capture the effect of experience in the game on the probability of cooperation. Since I directly control for cooperation decisions in this initial model, I do not expect time period fixed effects to have a significant effect over period-pooled OLS.

$$P_{it} = X_i + \beta_1 C_{it} + \beta_2 LT_{it} + \beta_3 O_{it} + \delta Z_{it} + \epsilon_{it}$$

Hypothesis one proposes that private and social incentives individually increase the marginal per capita return to cooperation. This marginal return to cooperation is captured above by the estimated coefficient β_1. Thus, in order to test this hypothesis, I need to decompose the estimated β_1 into the proportion that is governed by private incentives and the proportion that is dictated by social incentives. To do this, I construct two interaction variables which interact the binary cooperation variable with the separate incentives to produce the below regression.

$$P_{it} = X_i + \beta_1 C_{it} + \beta_2 LT_{it} + \beta_3 O_{it} + \beta_4 C_{it} \times LT_{it} + \beta_5 C_{it} \times O_{it} + \delta Z_{it} + \epsilon_{it}$$

Hypothesis one (a) predicts that the marginal per capita return to cooperation is increasing (decreasing) in private incentives (threat) which implies that β_4 is negative and significant. Similarly, hypothesis one (b) predicts that the marginal per capita return to cooperation is increasing in the social incentives which implies that β_5 is positive and significant.

5.4.2 Model two

Model two is used to test hypotheses two and three. Hypothesis two proposes that individuals are more likely to cooperate as the threat increases. Hypothesis three proposes an interaction between social and private incentives. Specifically, it proposes that there exists a component of the marginal

effect of social incentives that is independent of private incentives but there exists some interaction between private and social incentives which is positively correlated. In my model, I construct an interaction term between social and private incentives and test the estimated coefficient to be statistically significant from zero. The hypothesis proposes that as the relative threat increases, the marginal propensity to cooperate provided by social incentives decreases and as such I expect the coefficient of this interaction term to be negative.

The dependent variable for this experiment is the binary variable, C_{it} which states whether an individual cooperates in period t. The independent variable of interest is now the continuous measurement of the logarithmically transformed relative external threat, LT_{it}, and the continuous measurement of social incentives, O_{it}. In this experiment, I also proceed to add period fixed effects, X_t, since I now wish to control for the learning effect of cooperation within a set of repeated rounds. The baseline model is therefore the following.

$$C_{it} = X_i + X_t + \beta_1 LT_{it} + \beta_2 O_{it} + \delta Z_{it} + \epsilon_{it}$$

Hypothesis two can be tested from the baseline model in asserting that the estimated coefficient β_1 is statistically significant and negative such that as the relative threat increases, one's average propensity to cooperate diminishes.

A second regression includes the interaction term between social and private incentives as described below.

$$C_{it} = X_t + \beta_1 LT_{it} + \beta_2 O_{it} + \beta_3 O_{it} \times LT_{it} + \delta Z_{it} + \epsilon_{it}$$

Hypothesis three proposes that the estimated coefficient β_3 is significantly distinguishable from zero and negative. It also presupposes that the estimated coefficient β_2 is significantly greater than zero such that positive social incentives provide a positive marginal effect on an individual's propensity to cooperate in the absence of an external threat. If β_3 is negative, this demonstrates that the marginal effect of social incentives decreases as private incentives decrease.

5.4.3 Model three

Model 3 moves on to explore the impact of group similarity on the dynamic effects explored in model 2. To conduct experiment three, I begin by constructing a similarity composite index ζ_{it} which is the product of four separate similarity measures:

- The proportion of individuals in i's neighbourhood who are the same sex.
- One minus the proportional deviation of individual i's age from the neighbourhood average.
- The proportion of individuals in i's neighbourhood who come from the same social class using the three class system provided by Erikson and Goldthorpe (1992)).

Table 1: Summary Statistics of variables

	Unbalanced		Quasi-balanced	
	Mean	Standard dev	Mean	Standard dev
Progress	.875	.330	-	-
Cooperate	.690	.463	.730	.444
Log Threat	.356	.335	.355	.329
Proportion of other nbd members cooperating	.684	.362	.677	.361
Previous voted against	.027	.052	.027	.051
Previous neighbourhood residual success	.304	.257	.305	.251
Sex similarity of neighbourhood	.576	.179	.575	.186
Age similarity of neighbourhood	.784	.162	.774	.174
Proportion with lower socioeconomic status	.274	.244	.265	.242
Proportion with higher socioeconomic status	.274	.299	.296	.298
Proportion from same geographical division	.351	.202	.342	.200
Proportion from same geographical region	.431	.204	.415	.194
Number of total players	11.107	2.074	10.989	2.080
Merger occurred	.615	.487	.571	.495
Zeta	-	-	.119	.090
Observations	1261		703	

- The proportion of individuals in i's neighbourhood who come from the same geographical region in the United States as i. Here we use the four geographical regions as outlined by the United States Census Bureau (2021).

I expect that the effect an external threat has on an individual's incentive to cooperate via social norms increases with ζ_{it}. Thus I propose to include two additional covariate terms which interact ζ_{it} with O_{it} and $O_{it} \times LT_{it}$ to form the following regression:

$$C_{it} = X_t + \beta_1 T_{it} + \beta_2 O_{it} + \beta_3 O_{it} \times T_{it} + \beta_4 O_{it} \times \zeta_{it} + \beta_5 O_{it} \times T_{it} \times \zeta_{it} + \delta Z_{it} + \delta \epsilon_{it}$$

Hypothesis four proposes that the strength of the social norm effect implied by hypothesis three is greater the closer an individual is to the group average identity. I test this by constructing a composite measure of individual-group similarity, ζ_{it} and interacting this with the social-private incentive interaction from the previous regression. Hence, I expect to reject the null hypothesis that $\beta_5 = 0$ in favour of a positive coefficient.

6 Results

Table 1 displays the summary statistics of the primary variables from the unbalanced and quasi-balanced panel data sets. The former is used for model one while the latter for the remaining models. The statistics show that the mean and standard deviations are well preserved from the balancing procedure. Further, it is clear that there is a bias towards cooperation over both samples however this bias is augmented through the balancing procedure which is support for the use of a constrained sample in models two and three.

Table 2: Model 1

	(1) Baseline	(2) Interaction
Cooperate	0.138**	0.0220
	(0.0319)	(0.0451)
Log Threat	-0.160**	-0.161*
	(0.0466)	(0.0753)
Prop. of other nbd members cooperating	-0.0350	-0.137*
	(0.0423)	(0.0667)
Cooperate × Log Threat		-0.0118
		(0.0790)
Cooperate × prop. of other nbd members cooperating		0.200*
		(0.0793)
Constant	0.285	0.372*
	(0.174)	(0.173)
Controls	Yes	Yes
Period FE	No	No
Obs	1261	1261
Individuals	384	384
Adj R^2	0.4414	0.4456
IndFE	-0.254	-0.236
Sargan-Hansen stat	298.291	317.965
Sargan-Hansen p	0.0000	0.0000
Wooldridge test p	0.0115	0.0068

Cluster-robust standard errors in parentheses

* $p < 0.05$, ** $p < 0.005$

6.1 Model one

The estimated parameters to model one are summarised in table 2. This model investigates the effect that private and social incentives have on the attractiveness of cooperation in terms of enhancing an individual's probability of continuation in the game and thus one's expected payoff. Columns one and two represent the estimated coefficients of performing the regressions using one-way fixed effects analysis with cluster-robust standard errors in the parantheses. Both regressions were performed on a sample of 384 individuals over an average of 3.28 periods constituting a total sample of 1261 observations. The adjusted R squared value is computed using the 'areg' command in STATA to account for the loss of degrees of freedom due to fixed effects. The row 'IndFE' presents the correlation between the individual fixed effects and the independent variables. Since the random effects model assumes that this figure should be zero, a fixed effects model should see a non-zero number here. The row titled 'Sargan-Hansen stat' displays the test statistic from the cluster-robust modified Hausman test for fixed effects versus random effects. The row below displays the p-value that I use to reject the null hypothesis that a random effects model is a better fit for the model. The row titled 'Wooldridge

test p' displays the p-value that we use to reject the null hypothesis of no autocorrelation in the error terms in the panel set using the Wooldridge (2002) test via the command written by Drukker (2003).

Both the baseline and expanded model have a reasonable correlation figure between individual fixed effects and the explanatory variables and so I do not reject the use of these. Further, in both models, I reject the modified Hausman test null hypothesis that random effects is a better fit than fixed effects ($p < 0.001$ in both cases). Finally, the models do not reject the Wooldridge null hypothesis of no serial autcorrelation in the error terms at the significance level of 0.005 ($p = 0.0115$ and $p = 0.0068$) but both reject this null hypothesis at the significance level of 0.05. Regardless, I continue to employ cluster-robust standard errors to minimise the effect of heteroskedasticity on the estimated residuals.

The first column presents the coefficient estimates for the baseline model on the probability of progressing to the next stage of the game. This model rejects the null hypothesis that cooperation has no significant effect on the probability of progressing and thus expected utility at the 0.5% significance level. On average, the individual absolute marginal effect of cooperation is an increase in probability of progression by nearly 14 percentage points. This result suggests that on average cooperation is always beneficial to one's chance of survival regardless of private and social incentives.

The constructed and transformed variable for threat in the baseline model is also shown to be statistically significant and negative at the 0.5% significance level. Accordingly, an increase (decrease) in the relative threat (benefit) decreases one's probability of progressing to the next round regardless of one's decision to cooperate or the decisions of one's neighbourhood. This result suggests that the effect of private incentives on the marginal return to cooperation has an indirect component: an increased external threat crowds out the relative positive effect of cooperation on the probability of survival on average. Interestingly, the value of the external threat at which the benefit of cooperation is fully crowded out is 0.8625, in which case the size of the size of the external threat exceeds the size of one's own neighbourhood[13]. I herein refer to the value of the log threat at which the marginal implied utility benefit to cooperation is crowded out by the negative implied utility effect from an external threat as the **incentive compatibility threshold** as per Ackermann and Murphy (2019).

What is interesting in this baseline model is the non-rejection of the null hypothesis that the decision of others has no significant effect on one's probability of progression ($p = 0.409$). A minimum detectable effect size here given 80% power at a 0.005 (0.05) level of significance is 0.11844 (0.154) which is not unrealistic. Hence, I can say that I am powered to detect a realistic effect in this model and as such a lack of statistical power is likely not driving this null result. This interpretation suggests that if social incentives are to have an effect on one's expected utility, then these incentives may work directly through the marginal benefit of cooperation. Column 2 displays the estimated coefficients and their standard errors of the model which investigates this and the corresponding hypothesis.

The second estimated regression includes the interaction terms between the decision to cooperate and the private and social incentives. Hypothesis one proposes that the marginal effect of cooperation increases with private and social incentives. I investigate whether this relationship holds.

First, I look at social incentives. In this regression, the coefficient of the social incentive proxy

[13]This is computed since 0.8625 exceeds the value of the natural logarithm of 2: where the size of one's neighbourhood equals the size of the external threat.

in absence of individual cooperation is significant at the 5% level. Since this estimated coefficient is negative, there is suggestive evidence that when the individual does not cooperate and the proportion of individuals in the neighbourhood who do cooperate is nonzero, the probability that an individual progresses decreases from the baseline level. In other words, as the individual's decision to not cooperate lies further from the average decision of the neighbourhood, one may incur an expected utility loss due to social punishment.

In absence of social effects and external threat, the marginal benefit to cooperation is now not statistically distinguishable from zero ($p = 0.626$), Further, when the the external threat is non-zero (but the number of individuals cooperating is still zero) there is still an undetectable marginal benefit to cooperation. An F-test testing the significance of the sum of the coefficient of 'Cooperate' and 'Cooperate $\times$ Log Threat' cannot reject the null hypothesis that this linear combination of coefficient is equal to zero ($p = 0.9067$). Thus, I can interpret this result as in absence of any fellow cooperators, the marginal effect of one's decision to cooperation is negligible. However, a post-hoc power analysis suggests that this null result may be driven by a dearth of statistical power since minimum detectable effect sizes are perhaps a little larger than what would be considered reasonable.[14]

The marginal effect of cooperation with non-zero social incentives (when others in the neighbourhood are expected to cooperate) is in fact distinguishable from zero and positive at any level of private incentives. This is demonstrated through the positive and significant coefficient to the interaction variable between cooperation and the proportion of other individuals cooperating at the 5% level of significance ($p = 0.012$). A joint F-test on the linear combination of the coefficient estimates confirms that in the presence of a strong social norm to cooperate and of a mean value of the external threat, the marginal benefit of cooperation is significant at the 0.5% level ($p < 0.0001$). Combined with the above, this suggests that the marginal benefit from cooperation in the baseline model is directly dependent on the strength of the social incentive to cooperate: the more individuals that cooperate in one's neighbourhood, the greater the return to cooperation. This result therefore presents suggestive evidence for hypothesis one (b).

However, in combining the interaction effect of social incentives and the non-interaction effect of social incentives, one observes that the net effect of a social norm for cooperation appearing and subsequently adhering to it is not statistically distinguishable from zero with no social effects. This is shown by a joint F-test testing the linear combination of the coefficients for 'Prop. of other nbd members cooperating' and 'Cooperate $\times$ prop. of other nbd members cooperating'. The F-test cannot reject the null hypothesis that the linear combination of these effects is equal to zero ($p = 0.1768$). This suggests that cooperating is the best response when there is a social norm to do so but it does not produce a net increase in one's probability of survival on average.

Moving on to report the intuition behind private incentives, the above demonstrates that in absence of private incentives there is still a marginal return to cooperation given that a non-zero proportion of the individual's neighbourhood are also cooperating. However, since the coefficient of the interaction between private incentives and the decision to cooperate is indistinguishable from zero ($p = 0.882$), I cannot reject the null hypothesis that the direct marginal effect of cooperation

[14]The minimum detectable effect at 0.5% significance with 80% power for the marginal benefit of cooperating on the probability of survival in absence of social incentives and external threat is 17.36 percentage points. The minimum detectable effect at 0.5% significance with 80% power for the marginal benefit of cooperating on the probability of survival in absence of social incentives but with non-zero external threat is 30 percentage points.

Table 3: Model 2

	(1) Baseline	(2) Interaction
Log Threat	-0.164*	0.0528
	(0.0608)	(0.117)
Prop. of other nbd members cooperating	0.726**	0.848**
	(0.0640)	(0.0766)
Log Threat × prop. of other nbd members cooperating		-0.302*
		(0.139)
Constant	-0.0730	-0.276
	(0.396)	(0.409)
Controls	Yes	Yes
Period FE	Yes	Yes
Obs	699	699
Individuals	155	155
Adj R^2	0.4969	0.5015
IndFE	-0.235	-0.256
F_stat_p	0.00432	0.0198
Sargan-Hansen stat	46.805	48.950
Sargan-Hansen p	0.0006	0.0005
Wooldridge test p	0.2106	0.2211

Cluster-robust standard errors in parentheses

All regressions include controls

* $p < 0.05$, ** $p < 0.005$

is not affected by the private incentive to cooperate. In this model however, the coefficient to 'Log Threat' is still significant at the 5% level and as such an increase in the private threat is interpreted as crowding out the effect of cooperation. Nonetheless, this result does not support hypothesis one (a) since the direct marginal effect of cooperation is unaffected by the size of the external threat.

Result 1: Social effects significantly increase the marginal return to cooperation however the defined private incentives do not alter the marginal return. Rather an external threat crowds out the implied utility gains from cooperation.

In concern of collinearity between the private incentives and social incentives, I also conduct a variance inflation factor (VIF) test by regressing the private incentive proxy on the other explanatory variables in the baseline regression. I then compute the VIF test statistic by calculating the reciprocal of one minus the adjusted coefficient of determination of this regression. The adjusted coefficient of determination from this regression is 0.6359 resulting in a VIF test statistic of 2.746 which is below the rule of thumb threshold of 5 signalling that collinearity is not an issue in this model.

6.2 Model two

Table 3 presents the estimated independent variable coefficients, cluster-robust standard errors, regression and test statistics for the regressions performed for model two which are estimated using a two-way fixed effect method. Period fixed effects are also included to control for the learning mechanism that is intuitively hypothesised and observed in the literature (Fréchette and Yuksel, 2016). The added row 'F-stat p' presents the p-value for an F-test on the inclusion of period fixed effects. As per the a priori stipulation, I reject the null hypothesis that pooled OLS is best when this p-value is less than 0.005 however a p-value of less than 0.05 is sufficient to suggest that the intuition behind a learning mechanism may be present. Since experiment two employs the quasi-balanced panel, it is unsurprising to see a reduction in both the number of observations and the number of individuals or clusters that are present in the sample. Both models exhibit a moderate adjusted R squared and utilise the same vector of control variables. All models reject the period pooling of results at the 5% level of significance and so I continue to employ period-fixed effects. Further all models display a non-zero correlation between the individual fixed effects and the explanatory variables which indicates for the correct inclusion of individual fixed effects instead of random effects. This is additionally shown in the modified Hausman test which rejects the null hypothesis that random effects are more suitable than fixed effects for all three models. Finally, Wooldridge tests for serial autocorrelation in the error term are performed which cannot reject the null hypothesis of no serial autocorrelation. However, just like in experiment one, I continue to employ cluster-robust standard errors to control for heteroskedasticity.

Column one presents the results to the baseline regression of model two which constructs a two-way fixed effect model using the private incentive proxy (the natural logarithm of the relative external threat) and the social incentive proxy (the proportion of individuals in the individual's neighbourhood who cooperate in time t, not counting the individual) as the key independent variables only. As predicted by the theoretical model and from the previous subsection, this baseline model shows a significant positive relationship on average between social incentives and the propensity to cooperate ($p < 0.001$) and a significant negative relationship on average between the relative external threat and the probability of cooperating although only at the 5% significance level. This provides suggestive evidence for hypothesis two.

Result 2: On average, an increase in external threat decreases the probability that an individual cooperates when controlling for the social norm.

Column two presents the results of the interaction regression which additionally includes a variable for the interaction between the private and social incentives. In this regression, I am able to distinguish between the cases where private incentives are zero and social incentives are zero. First, I observe that in absence of social incentives, the effect of private incentives found in the baseline regression is not distinguishable from zero ($p = 0.653$). However, this null result should be interpreted with caution since a post-hoc power analysis shows that the minimum detectable effect of private incentives in absence of neighbourhood cooperation is unrealistically large compared to the baseline model (MDE=0.45045 at 0.5% level of significance). From this result, I cannot suggest that in absence of the possibility to conditionally cooperate, one is incentivised to unconditionally cooperate at any level of external threat.

To test hypothesis three, I must test the coefficients to social incentive parameters. First, I observe the marginal effect of social incentives in absence of an external threat. Table three shows

that the estimated coefficient to the proportion of other neighbourhood members cooperating is statistically significant at the 0.5% level of significance and positive. This suggests that even in the occasion of zero external threat, positive social incentives increase one's propensity to cooperate.

In the case of a non-zero external threat, a significantly negative coefficient of the interaction variable at the 5% level of significance provides suggestive evidence that an increasing external threat diminishes the marginal effect of social incentives on one's probability of cooperation. This provides suggestive evidence that the effect identified in the baseline model and result two occurs through an average decrease in one's propensity to conditionally cooperate only and as such provides suggestive evidence for hypothesis three. This effect suggests that individuals are more likely to defect from the social norm as the size of the external threat increases.

Result 3: In the absence of external threats, social incentives increases one's propensity to conditionally cooperate. This effect diminishes as the external threat grows.

A variance inflation factor test was performed for this experiment by regressing the private incentives on the remaining independent variables minus the interaction effect. The VIF test statistic in this case is 2.21 which is below the common threshold of five which indicates that multicollinearity is not a problematic issue in this regression.

6.3 Model three

Table 4 presents the regression results to model three alongside regression two from model two in column 1. In observing column two, I have included two additional interaction terms which interact the social incentives and the social-private incentive interaction with the similarity composite index - zeta (ζ). Summary statistics on zeta from table 1 show that the mean of this variable is very low considering it is computed to range between zero and one. Further, any inference drawn from interactions with zeta are context specific to the formation of zeta and different results may be drawn from different composites.

In column two of table 4, I see that the regression displays similar test statistics to column one indicating that the model rejects the same null hypotheses of statistical tests as in experiment two. Of interest, are the estimated coefficients to the two new interaction terms presented in column 2. In fact, neither coefficients of the new additional terms are statistically distinguishable from zero at the 5% level of significance and the coefficients which persist in both regressions in table four do not change in their significance and are not statistically distinguishable from each other at the 5% level of significance. This suggests that my formulation of zeta holds no influence on the size of the social incentive effect in this specific context and as such I do not find evidence to support hypothesis four.

A post-hoc power analysis to compute the minimum detectable effect finds that I am not powered at 80% to identify effects of an expected size however. Consider the interaction between the similarity index and the proportion of individuals cooperating in one's neighbourhood. The estimated standard error of this variable is 0.574 giving a minimum detectable effect of 1.607 (2.21) at the 5% (0.5%) level of significance. Hence, this null effect may be due to a lack of power.

In light of this null result, I considered a number of robustness tests to ascertain whether this inability to find a significant effect was due to the composition of the zeta index. Accordingly table 5 presents the regression results to four separate regressions which consider each element of the zeta

Table 4: Model 3

	(1) Interaction	(2) With zeta
Log Threat	0.0528	0.0428
	(0.117)	(0.116)
Prop. of other nbd members cooperating	0.848**	0.952**
	(0.0766)	(0.109)
Log Threat × prop. of other nbd members cooperating	-0.302*	-0.381*
	(0.139)	(0.162)
Prop. of other nbd members cooperating × Zeta		-0.869
		(0.574)
Log Threat × prop. of other nbd members cooperating × Zeta		0.702
		(0.543)
Constant	-0.276	-0.364
	(0.409)	(0.412)
Controls	Yes	Yes
Period FE	Yes	Yes
Obs	699	699
Individuals	155	155
Adj R^2	0.5015	0.5017
IndFE	-0.256	-0.260
F_stat_p	0.0198	0.0247
Sargan-Hansen stat	48.950	52.930
Sargan-Hansen p	0.0005	0.0004
Wooldridge test p	0.2211	0.2073

Cluster-robust standard errors in parentheses

All regressions include controls

* $p < 0.05$, ** $p < 0.005$

composite separately. These elements are the proportion of individuals in the the neighbourhood who are the same sex as individual i, the relative deviation of i's age from the neighbourhood average, and the proportion of subjects in i's neighbourhood who have a higher or lower social economic status to the individual i according to the 5 class system outlined by Erikson and Goldthorpe (1992).

Table five shows that in fact none of these elements have a significant interaction effect with social incentives when modelled individually. Standard errors in these models are much lower than in the previous estimation which represent smaller minimum detectable effect sizes and thus a smaller chance of the null result being a consequence of low power.

Result 4: There is no identified effect of group similarity on the marginal effect of social incentives to cooperate.

Table 5: Model 3 robustness tests

	(1) Sex	(2) Age	(3) Lower SES†	(4) Higher SES
Log Threat	0.224	-0.215	0.101	0.0117
	(0.171)	(0.229)	(0.0846)	(0.121)
Prop. of other nbd members cooperating	1.041**	1.060**	0.82**	0.861**
	(0.183)	(0.167)	(0.0846)	(0.100)
Log Threat × prop. of other nbd members cooperating	-0.273	-0.316*	-0.287*	-0.313*
	(0.118)	(0.150)	(0.137)	(0.143)
Prop. of other nbd members cooperating × Zeta element	-0.369	-0.272	0.0845	-0.0427
	(0.302)	(0.196)	(0.150)	(0.192)
Log Threat × prop. of other nbd members cooperating × Zeta element	-0.266	0.357	-0.250	0.203
	(0.239)	(0.279)	(0.191)	(0.216)
Constant	-0.490	-0.344	-0.297	-0.263
	(0.416)	(0.413)	(0.415)	(0.410)
Controls	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes
Obs	699	699	699	699
Individuals	155	155	155	155
Adj R^2	0.5043	0.5015	0.5019	0.5006
IndFE	-0.267	-0.264	-0.271	-0.245
F_stat_p	0.0135	0.0164	0.0128	0.0199
Sargan-Hansen stat	55.148	49.555	55.207	47.345
Sargan-Hansen p	0.0002	0.0011	0.0002	0.0020
Wooldridge test p	0.2292	0.2091	0.2320	0.2049

† SES=Socioeconomic status according to the table by Erikson and Goldthorpe (1992)

Cluster-robust standard errors in parentheses

All regressions include controls

* $p < 0.05$, ** $p < 0.005$

7 Robustness tests

7.1 Is the effect of private incentives continuous and linear?

The results of the previous section are subject to the theoretical and empirical construction of the proxy variables. In particular, the private incentive proxy variable is not standard according to the literature since the natural logarithmic transformation of the external threat component assumes a single continuous linear marginal effect and doesn't account for potential structural breaks. It may be argued that for low values of external threat such that cooperating is still incentive compatible according to model one, an increasing threat would incentivise individuals to cooperate since the environment is becoming more hostile (Bowles and Gintis, 2011). One need only reflect on results of third-party punishment games to observe evidence of individuals acting more pro-socially when there is a risk of being punished by an observer if they do not (Fischbacher et al., 2001). In this study, the external threat is not defined by an altruistic observer but instead as a constant exogenous threat to a society. Once an external threat crowds out the benefit of cooperation (in this case that the size of the rival alliance is not smaller than the size of individual i's neighbourhood), the positive influence that a threat may have on individual's propensity to cooperate may disappear.

Thus to test for the presence of a structural break under this intuition, I reproduce model two once more and include an additional variable which measures the value of 'Log Threat' when the threat is sufficiently high that the marginal benefit of cooperation is crowded out, and zero otherwise. The threshold value is computed according to coefficient estimates of the baseline regression in model one and is equal to 0.8625. This added variable is included alone and is interacted with the social incentive proxy in column two.

The results of this model are displayed in table 6. First, I conduct an F-test to test for a gain in goodness of fit by including these two new variables. I obtain a test statistic of 1.25 and 2.249 for the baseline and interaction regressions respectively. Hence, I cannot reject the null hypothesis that the models presented in table six have superior goodness of fit over those presented in table three at the 5% level of significance. Further, I conduct a t-test for the significance of the new variable in the baseline model which does not reject the null hypothesis that the coefficient of the 'Log Threat (High)' variable is equal to zero (p=0.228). Further, I conduct an F-test for the inclusion of the new variables for high threat and find that I cannot reject the null hypothesis that both the coefficient of 'Log Threat (High)' and the coefficient of 'Log Threat (High) $\times$ prop. of other nbd members cooperating' are equal to zero (p=0.1683). Therefore, this model rejects the presence of a structural break at the specified cutoff as predicted from model one.

Nonetheless, I am suspicious of this result for two reasons. First, it is suggested that in the baseline regression, there exists a significant effect of the external threat on individual propensity to consume when the threat is high. This is computed through a joint F-test on the significance of the linear combination of the two coefficients for this regression (p=0.0062). Second, it is interesting to note that in the interaction regression, the effect of the interaction between external threat and social incentives is statistically significant at the 5% level but this effect becomes just statistically insignificant (p=0.0774) when the threat is high. Thus, I am skeptical that perhaps the non-rejection of the null of no structural break is sensitive to the prescribed cutoff point. The cutoff point may in fact differ from the point of incentive capability for two reasons: risk and the psychological cost to cooperation. The former may come into play if one considers that those in their current alliance may deviate from the voting rules or if one considers that those in a rival alliance may not all co-

Table 6: Robustness test 1

	(1) Baseline	(2) Interaction
Log Threat	-0.0783	0.405
	(0.0863)	(0.218)
Log Threat (High)	-0.0919	-0.413
	(0.0759)	(0.218)
Prop. of other nbd members cooperating	0.715**	0.882**
	(0.0654)	(0.078)
Log Threat × prop. of other nbd members cooperating		-0.665*
		(0.244)
Log Threat (High) × prop. of other nbd members cooperating		0.424
		(0.235)
Constant	-0.103	-0.271
	(0.402)	(0.409)
Controls	Yes	Yes
Period FE	Yes	Yes
Obs	699	699
Individuals	155	155
Adj R^2	0.4971	0.5039
IndFE	-0.2519	-0.2472
F_stat_p	0.0077	0.0223
Sargan-Hansen stat	51.417	47.468
Sargan-Hansen p	0.0002	0.0020
Wooldridge test p	0.2041	0.2155

Cluster-robust standard errors in parentheses

All regressions include controls

* $p < 0.05$, ** $p < 0.005$

operate with their voting rules. Thus if risk preferences do play a role, I may have overestimated or underestimated the cutoff point at which a threat is interpreted as high depending on which risk dominates. The latter is difficult to quantify in this study and is often monetised in experiments by the 'temptation cost' as termed by Mengel (2018).

To account for these underlying mechanisms and the potential that the relationship may not be linear as assumed by the previous models, it may instead be more appropriate to consider a non-linear relationship between private incentives in a natural logarithm form and the propensity to cooperate.[15] This would allow for continuity along the threshold itself and so does not require the assumption of a specific cutoff value of external threat at which cooperation is no longer privately an equilibrium, however it does permit an evolving marginal effect of private incentives. Table 7

[15]I do not specify a functional form for the distribution of the errors in the theoretical model, thus a non-linear consideration is not in violation of the theoretical predictions of the relationship.

Table 7: Robustness test 2

	(1) Baseline	(2) Interaction
Log Threat	-0.0788	0.720*
	(0.147)	(0.331)
Log Threat^2	-0.0902	-0.734*
	(0.138)	(0.341)
Prop. of other nbd members cooperating	0.722**	0.914**
	(0.0644)	(0.0765)
Log Threat × prop. of other nbd members cooperating		-1.103**
		(0.368)
Log Threat^2 × prop. of other nbd members cooperating		0.864*
		(0.351)
Constant	-0.100	-0.330
	(0.406)	(0.413)
Controls	Yes	Yes
Period FE	Yes	Yes
Obs	699	699
Individuals	155	155
Adj R^2	0.4963	0.5048
IndFE	-0.244	-0.269
F_stat_p	0.00803	0.0337
Sargan-Hansen stat	50.113	50.481
Sargan-Hansen p	0.0004	0.0008
Wooldridge test p	0.2089	0.2335

Cluster-robust standard errors in parentheses

All regressions include controls

* $p < 0.05$, ** $p < 0.005$

displays the results of this robustness test which includes the added variable Log Threat^2 which is the square of the natural logarithm of one plus the relative threat value.[16]

Once more, I estimate the significance of the new unrestricted model using a standard F-test. I obtain test statistics of 0.380 and 2.661 for the baseline and interaction regressions respectively. This again indicates that I cannot reject the null hypothesis that the regressions in table seven are superior in goods of fit with respect the regressions in table three. Further, a t-test on the significance of the estimated coefficient to Log Threat^2 does not reject the null hypothesis that the coefficient is equal to zero. However a joint F-test suggests that the linear combination of the two coefficients is statistically significant from zero (p=0.0058). The joint F-test for the inclusion of the

[16]While it may be phrased such that this model was considered post hoc, the potential for such a problem involving the cutoff point was identified before any such regressions were ran and the use of a quadratic variable was prescribed according to the pre-analysis plan.

two new variables in the interaction column does reject the joint null hypothesis that both estimated coefficients are equal to zero at a 5% level of significance, albeit only just (p=0.0498). Thus, it appears that the inclusion of a quadratic variable may be suitable.

In the baseline model, the coefficients for the linear and quadratic terms of the private incentive are not statistically significant individually. However, as aforementioned, the combined marginal effect of Log Threat on the propensity to cooperate in the baseline model is statistically significant at the 5% level (p=0.0058). This proposes that there is a decreasingly negative and significant effect of an external threat on the propensity to cooperate which explains why I was unable to find a significant effect in the previous robustness test's baseline model of low values of threat on the propensity to cooperate: the marginal effect identified of external threat is now level-specific and is thus negligible at small levels of threat but increases in magnitude with the size of the external threat. This should, however be interpreted with caution due to the large standard errors present in this regression estimation. One should rightly be concerned of power issues in identifying this effect in the baseline regression. Specifically, I am not powered to identify the small effects that are implied for low values of log threat.

Regression two presents a similar story when the interaction effects are included. If one begins with interpreting the results with a zero value for social incentives (the top two coefficient estimations in isolation), then the model provides suggestive evidence that at very low but positive private incentives, an increase in threat does in fact provide a net positive marginal effect on the propensity to cooperate since both coefficients to the private incentives are significant at the 5% level. This can be interpreted as a suggestion that low but increasing threats incentivise unconditional cooperation on average. As the threat increases, the marginal effect of an external threat on the propensity to cooperate will eventually be negative since the coefficient to the quadratic term is negative and significant at the 5% level of significance. This may explain why one observed an average null effect of private incentives on propensity to cooperate in absence of social incentives in regression two of the main model two.

One can observe that even in a situation with zero outside threat, individuals will cooperate if sufficient others are cooperating. This is shown by the statistically significant and positive coefficient to the proportion of other neighbourhood members cooperating.

The estimated coefficient of the linear interaction term is negative and significant at the 0.5% level while the estimated coefficient of the quadratic interaction term is positive and significant at the 5% level. This interestingly suggests that there exists some minimum value of the marginal effect of social incentives on the propensity to cooperate as external threats increase. One can compute that the value of the relative external threat that corresponds to this minimum value is 0.693.

A number of observations are worthy of mention from this result. First, it is evident that one's sensitivity to a social norm is in fact unstable according to the level of private incentives: the external threat. I interpret this as evidence that one's incentive to cooperate according to a norm evolves as one's private incentive to cooperate varies. This leads to the second observation: the total effect of private incentives on one's inclination to cooperate is dependent on the size of the social norm. Under weak social norms of cooperation, unconditional cooperation in low threat scenarios can increase one's propensity to cooperate. However, under strong social norms of cooperation, the estimated coefficients suggest that there is little variation of one's propensity to cooperate as the external threat evolves. Nonetheless, if one expects a social norm to break down as the external

threat grows large, then the unconditional cooperation effect dominates and cooperation may break down à la Kandori (1992).

Finally, one can use this regression to test the identification of the incentive compatibility threshold as the point at which an individual alters their strategy dynamics by computing the value of the Log Threat at which the direct and indirect marginal effect of the external threat changes sign. One's propensity to unconditionally cooperate reaches a maximum value at the point where the private incenive proxy is equal to 0.693. This value suggests that individuals rationally respond to their private incentives since 0.693 lies close to true value of the transformed private incentive corresponding to equal sizes of one's own neighbourhood and the outside threat which is the theoretically predicted incentive compatibility threshold.[17]

Result 5: When the external threat is low and cooperation is the private equilibrium, individuals become less sensitive to social pressures to cooperate as the threat increases and instead do so unconditionally. When the external threat is higher, individuals favour conditional cooperation, only cooperating if a norm is present for doing so.

7.2 Is the marginal effect of social incentives robust to the specification of the proxy?

In composing the proxy for social incentives, I implicitly assume that open and unrestricted communication allows individuals to correctly identify the contemporaneous social norm within their neighbourhood. However, Realpe-Gómez et al. (2019) argue that a signal of cooperation which is visible and comprehensible may not have the same effect as the number of individuals who actually cooperate since there is always a risk of unilateral defection. I provide evidence that a degree of risk aversion is present in the previous section. In this robustness test, I relax the assumption that individuals form risk neutral perfect estimates of the proportion of individuals to cooperate and instead assume that the social norm is contemporaneously estimated by the more visible measure of the proportion of individuals in the neighbourhood who are members of the alliance. In other words, the social norm is dictated by the proportion of individuals who signal their intention of cooperating and not the proportion of individuals who in fact do cooperate. Thus, this secondary proxy for social incentives will sometimes overestimate the proportion of individuals within a neighbourhood who cooperate. It is noteworthy that the regressions performed in this subsection were in fact designed after the analysis of the previous results and so any interpretations drawn from these experiments should be made extremely cautiously.

The new social incentive proxy, O_2, is skewed slightly more to the right than the first proxy O with a mean of 0.715 against the original 0.684. This reflects the fact that individuals may defect even if they are a member of an alliance and so the secondary proxy O_2 is not a perfect measure of the social norm. However, O_2 may better reflect individual social incentives to cooperate since it is more visible to players. I conduct a series of regressions using this secondary social incentive proxy and compare the results with the previous main models.

Table 8 displays the results to model one's interaction regression using each social incentive proxy.

[17] A t-test testing the distinction of this value of private incentives from the point at which the rival alliance is equal to the size of one's neighbourhood cannot reject the null hypothesis that this difference is equal to zero ($p = 0.1372$).

Table 8: Robustness test 3

	(1) O_{it}	(2) O_{2it}
Cooperate	0.0220	0.0400
	(0.0451)	(0.0459)
Log Threat	-0.161*	-0.168*
	(0.0753)	(0.0744)
Social incentive proxy	-0.137*	-0.144*
	(0.0667)	(0.0566)
Cooperate × Log Threat	-0.0118	-0.00196
	(0.0790)	(0.0770)
Cooperate × social incentive proxy	0.200*	0.167*
	(0.0793)	(0.0719)
Constant	0.372*	0.378*
	(0.173)	(0.174)
Controls	Yes	Yes
Period FE	No	No
Obs	1261	1261
Individuals	384	384
Adj R^2	0.4456	0.4465
IndFE	-0.236	-0.243
Sargan-Hansen stat	317.965	293.986
Sargan-Hansen p	0.0000	0.0000
Wooldridge test p	0.0068	0.0107

Cluster-robust standard errors in parentheses

* $p < 0.05$, ** $p < 0.005$

The first column is a replication of the interaction regression from table 2 while the second column displays the results of the same regression but with the use of the secondary social incentive proxy, O_2. In comparing the results of the two regressions, there are no differences in the results of t-tests when using the secondary social incentive proxy. Therefore, it can be inferred that the results and intuition from experiment one are robust to the choice of social incentive proxy. This is not surprising since model one does not measure individual behavioural decisions as a dependent variable: rather it computes the marginal utility benefit of cooperating. Thus, the robustness of this result to the specification of the social incentive proxy suggests that the secondary social incentive proxy is a good signal for those who will eventually cooperate.

Table 9 displays the results to the linear interaction regression in model two using each social incentive proxy in columns one and two. Columns three and four then present the results to the quadratic interaction regression from table 7 (robustness test 2) using each social incentive proxy. In comparing columns one and two, there is evidence that the linear interaction effect suggested in model two is not robust to the use of social incentive proxy since the estimated marginal effect of

Table 9: Robustness test 4

	Linear model		Quadratic model	
	(1)	(2)	(3)	(4)
	O_{it}	O_{2it}	O_{it}	O_{2it}
Log Threat	0.0528	0.00451	0.720*	1.001**
	(0.117)	(0.130)	(0.331)	(0.330)
Log Threat^2			-0.734*	-1.141**
			(0.341)	(0.332)
Prop. of other nbd members cooperating	0.848**	0.709**	0.914**	0.768**
	(0.0766)	(0.0900)	(0.0765)	(0.0859)
Log Threat × prop. of other nbd members cooperating	-0.302*	-0.161	-1.103**	-1.157**
	(0.139)	(0.148)	(0.368)	(0.367)
Log Threat^2 × prop. of other nbd members cooperating			0.864*	1.125**
			(0.351)	(0.350)
Constant	-0.276	-0.371	-0.330	-0.386
	(0.409)	(0.457)	(0.413)	(0.449)
Controls	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes
Obs	699	699	699	699
Individuals	155	155	155	155
Adj R^2	0.5015	0.4584	0.5048	0.4669
IndFE	-0.256	-0.2470	-0.269	-0.2518
F_stat_p	0.0198	0.0003	0.0337	0.0016
Sargan-Hansen stat	48.950	55.288	50.481	58.703
Sargan-Hansen p	0.0005	0.0001	0.0008	0.0001
Wooldridge test p	0.2211	0.0694	0.2335	0.0919

Cluster-robust standard errors in parentheses

All regressions include controls

* $p < 0.05$, ** $p < 0.005$

the interaction between private and social incentives is no longer statistically distinguishable from zero at the 5% level of significance in using the secondary social incentive proxy. This suggests that the incentive to conditionally cooperate is not sensitive to the number of individuals in an alliance but is sensitive to the number of individuals who cooperate in the linear functional form. Thus one can infer that individuals are able to use open and unrestricted communication to infer who in an alliance is likely to cooperate and who is not.

Turning to the non-linear effect that is explored in the previous section, the results and intuition of this model are in fact robust to the use of social incentive proxy since every estimated coefficient of interest except the constant is significant at the 0.5% level in the robustness test. This provides further evidence that the effect of the external threat is not linear in this model and that the marginal effect of private incentives differs depending on whether cooperation is a private equilibrium or not

according to the size of the external threat.

8 Discussion

This thesis has investigated the stability of strategic cooperation under varying private and social incentives and very high stakes. Private incentives refer to the individual private return to cooperation and is proxied for by the existence of an external threat. Social incentives refer to the presence of a social norm for cooperation within one's neighbourhood. The results demonstrate that while social incentives dictate the marginal utility benefit to cooperation, an external threat may crowd out this effect. Nonetheless, I show that individuals do vary in their willingness to conditionally cooperate as the perceived private return to cooperation varies. When cooperation is an equilibrium strategy, a growing threat represents a shift from conditional cooperation towards unconditional cooperation. However, when cooperation is no longer equilibrium and the threat fully crowds out the marginal utility benefit to cooperation, an increase in this threat represents a strategy shift towards conditional cooperation. Finally, I do not find evidence that this instability is sensitive to group composition of similar identities.

Result one demonstrates that individuals receive a higher marginal benefit from cooperation the greater the proportion of one's group cooperating. The results from model one also suggest that individuals receive a 'punishment' in terms of a decrease in their likelihood of survival if they do not cooperate and there is a norm to cooperate. This punishment increases in magnitude the stronger the norm. One can interpret this as further evidence for the results of Fehr and Gächter (2000*b*) since punishment for norm violation is shown to increase as a greater proportion of individuals adhere to the norm. Even if there is an imperfect norm for cooperation, individual social punishment may be able to convert non-instrumental reciprocators. The results also demonstrate that the greater the proportion of individuals cooperating, the greater the average return to cooperation.

The first result of this thesis additionally proposes that external threats crowd out the benefit of cooperation but do not directly affect the marginal return to cooperation in this specific setting. This suggests that the theoretical model requires refinement to account for this specific form of private incentives. Thus, it is plausible to infer that despite cooperation having a positive impact on one's probability of survival given some proportion of others cooperating, it may not be optimal for one to cooperate if cooperation is costly and the threat is too great. I refer to this as incentive incompatibility. This provides some empirical justification for the equilibrium condition of Brock and Durlauf (2001) and suggests that to maintain the attraction of a cooperative signal, it is worthy to inform individuals of the private value of cooperating in face of a danger. This result also implies that with weaker social norms for cooperation, it requires a smaller external threat to crowd out the benefit to cooperation, and therefore highlights the value in encouraging coordination on cooperative strategies to achieve higher cooperation.

Result two builds on the above by showing that for any social norm, an external threat diminishes one's incentive to cooperate on average. Result three further shows that individuals on average grow increasingly less sensitive to the social norm as mutual cooperation becomes less lucrative. I therefore propose that conditional cooperation becomes less likely in this case, which provides a dynamic context to the effect found by Dreber et al. (2014) and Arechar et al. (2018). As individuals become increasingly aware of their risk of leaving the game, their propensity to employ instrumental reciprocity diminishes. This result is widely supported by the literature (see Embrey et al. (2018);

Reuben and Suetens (2011)). An explanation for this result is that individuals are less sensitive to norms since they hold less belief in the realisation of any future punishment since others may also be incentivised to defect (Balliet et al., 2016). This dynamic shift away from instrumental reciprocity is reinforced if one believes that others will behave in the same way and thus conditional cooperation breaks down (Kandori, 1992).

A dynamic study of the mechanism behind this effect is summarised in result five, which outlines the proof of unstable phenotypes from this thesis. Unlike the results of Dreber et al. (2014) and Arechar et al. (2018), I find an interdependent relationship between social and private incentives such that the emergence of conditional cooperation strategies are jointly determined by norms and private payoffs. I begin with a discussion on my results for the instance where cooperation is incentive compatible and follow with a discussion on norm compliance under high threats. First, I find that individuals transfer from a conditional cooperation strategy towards unconditional cooperation as external threat increases but cooperation is still incentive compatible. As such, under weaker norms of cooperation, individual probability of cooperation actually increases on average with an increasing external threat. On the other hand, under strong social norms of cooperation, an increasing external threat causes individual propensity to cooperate to fall on average. I propose explanations for each effect below.

First, under weak social norms of cooperation and an increasing but small external threat, a rational but non-cooperating individual may acknowledge the imminent threat to their survival and the growing value in encouraging cooperation. Thus, forward looking individuals may begin to cooperate as the threat rises in hope of establishing a stronger norm in the next period. Unfortunately, my model does not incorporate such inter-temporal forward looking behaviour since I assume that individuals are perfectly able to predict the contemporaneous cooperative decisions of others through communication (Realpe-Gómez et al., 2019). Therefore, there is no possibility for dynamic feedback to reinforce any degree of instrumental reciprocity, which is especially important if the visibility of others' cooperative actions is not perfect in the current period since beliefs may not sufficiently overlap with reality (Irlenbusch et al., 2018).

If this explanation holds, then my results would support the interpretation of Nagatsu et al. (2018), who found that a group of free-riders may shift to cooperative strategies when they are aware of the homogeneity of the groups' intentions and the opportunity cost of mutual defection. Moreover, an additional condition of their work is that individuals understand that there is a social expectation to cooperate among the free riders. My results indicate that individuals grow less sensitive to a social norm when threat is low but increasing. However, I measure a social norm as a realised observation and not in terms of forward looking beliefs which is implied by Nagatsu et al. (2018). In particular, they propose that forward looking individuals are aware that individuals in this case believe in the value of cooperation and that this is common knowledge which does not directly contradict my model since I assume that cooperation is known to be beneficial to one's probability of survival. This line of reasoning also supports the hypothesis that pro-social individuals may instigate a dynamic social norm with patience by cooperating in the current period and hoping that this norm will be adopted by others (Kurokawa et al., 2018; Haag and Lagunoff, 2007).

The other side of this discussion is that under strong social norms for cooperation, individual propensity to cooperate on average falls as the external threat level rises but remains below the incentive compatibility threshold. A potential explanation for this effect is that individuals are in fact imperfectly conditionally contributing (Fischbacher et al., 2001). Imperfect conditional coop-

eration implies that as the benefit to mutual cooperation is crowded out by an external threat, individuals expect that the social norm may be broken. Since individuals in this case would believe others may be more likely to defect, they too would defect in order to avoid being left with a 'sucker payoff'. Nonetheless, this effect is still in line with that of Dreber et al. (2014), Arechar et al. (2018), and Embrey et al. (2018) since it predicts that individuals defect from the conditional cooperation strategy as it becomes less beneficial to engage in this activity.

Irrespective of the above dynamic effects, when one observes the point estimates of coefficients in model two, one can see that since social incentives have a direct positive impact on the probability of cooperation, regardless of the external threat presented, one will always have a higher probability of cooperation under a strong social norm than under a weak social norm at any level of the external threat. This implies that policies should perhaps focus on exploiting the mechanism of social coordination before turning to the framing of an external threat. Further, these two types of policies should not be considered in isolation since their dynamic impact is interdependent. A promising interpretation of this is that since individuals are privately incentivised to cooperate even without an explicit norm, then an instrumental forward looking cooperation strategy is evolutionary stable since defection that weakens the social norm will merely lead to the more patient individuals continuing to cooperate and wait out the lull until a return to conditional cooperation and thus a strong social norm is re-adopted (Traxler and Spichtig, 2010; Dal Bó and Fréchette, 2018).

When cooperation is not incentive compatible, individuals in fact continue to exhibit varying degrees of reciprocity depending on expected social norms and private incentives. Result three demonstrates that on average, as the threat continues to increase beyond the incentive compatibility constraint, individuals are less likely to cooperate on average. Result five then disentangles the specific mechanism of this by demonstrating that in fact individuals do become more sensitive to social norms as the threat increases past the constraint but are less likely to unconditionally cooperate. Thus once more, in this domain, the effect that a external threat has on one's propensity to cooperate is dependent on the strength of the social norm as per Realpe-Gómez et al. (2019).

In face of a strong social norm, individuals' propensity to cooperate is dynamically little affected. However, as the social norm to cooperate becomes weaker, individuals are less likely to cooperate which suggests that the social norm and belief in such is essential to maintain this cooperation decision when private incentives fall short. Thus while individuals grow increasingly sensitive to the social norm as the threat increases, if the average cooperation rate is expected to decrease with an increasing threat, then my results suggest that cooperation will break down.

My results above implicitly support the theory that individuals shift from conditional cooperation to defection strategies when the incentive compatibility constraint does not hold and the external threat increases (Dreber et al., 2014; Arechar et al., 2018). Intuitively, this would be since individuals are less likely to cooperate unconditionally and are simultaneously imperfect conditional cooperators. As such individuals growing more sensitive to the believed average cooperation decision of the group expect that this average decision will fall with increasing threat which results in a breakdown of cooperation. My results also suggest that this is not an inevitable outcome: if a social norm for cooperation is continuously believed in and this social norm is able to withstand the heightened level of threat, then individuals will continue to cooperate since they grow increasingly attune to the expected average decision of the group. This is likely due to the increased danger of being outside of an alliance should punishment arise in the next round. I do not identify a specific mechanism that may facilitate social norm maintenance, however the presence of sufficient pro-social

forward-looking cooperators may in fact increase beliefs on the average cooperative decision of the group as suggested by Dreber et al. (2014).

Notably, my results do not prescribe that the strategy of always defection is increasingly adopted by all in this state. Instead, I propose that individuals grow increasingly sensitive to the social norm or the average cooperation decision of the group and thus a decision to defect may not be as selfish as other researchers imply (Arechar et al., 2018). Instead, one may be continuing to adapt to one's milieu by conforming to the expected average cooperative decision. This mirrors the conjecture by Kandori (1992) that in the presence of conditional cooperators, a negative shock to the expected average cooperation decision can result in the unraveling of a cooperative equilibrium. Thus to extend the interpretation of Dreber et al. (2014), I propose that the presence of pro-social unconditional cooperators who do not respond to this unravelling will provide a lower limit to this expected average cooperation decision and sufficient numbers of these will enable an equilibrium of mutual cooperation to prevail under high threat.

Thus, on the one hand, my results do support the proposition that pro-social unconditional cooperators are key to maintaining cooperation when the act is not incentive compatible from an individual strategy perspective as per Dreber et al. (2014). However, I additionally propose that mutual cooperation may be maintained if one can reinforce the believed expected average cooperative decision through transparency of decisions Irlenbusch et al. (2018).

While this result is important to the myopic policy maker, one limitation of my interpretation is that I am unable to identify any effect of leniency which may itself provide a mechanism for maintaining a cooperative equilibrium. This is because a two-way panel design is implemented which estimates coefficients as averages over time. Further, I do not include inter-temporal independent variables which may show leniency. Thus, it may be that under imperfect monitoring, individuals exhibit leniency and forgiveness in favour of a cooperative equilibrium (Dal Bó and Fréchette, 2018). This would particularly be the case if individuals have grown accustomed to being in a state of cooperation and will thus be drawn to leniency by status-quo bias (Kahneman et al., 1991). Moreover, this will be efficiency improving, particularly if the unravelling of mutual cooperation is triggered not by a private decision to defect, but by a stochastic error (Traxler and Spichtig, 2010), or is triggered by selfish myopic anomalies (Balliet et al., 2016; Proto et al., 2019).

Surprisingly, result four states that there is no identified effect of group similarity on individual propensity to conditionally cooperate, which implies that individuals place equal trust (or distrust) on their co-players regardless of their identities. One may reason for this null result if one considers the neighbourhoods as a sub-game identity, thus this demonstrates how identities can be endogenously prescribed by experimenters (Chen et al., 2014). Further, since individuals may not be randomly assigned to initial tribes and can self-sort into neighbourhoods should they wish prior to the beginning of my data collection period, identifying factors may not be useful in predicting individual norm sensitivity. To further investigate this, one should randomly assign group membership to observe the influence of social identifying factors and group similarity on individual reciprocity. Interestingly hence, my results suggest that as long as neighbourhoods are defined and members share a common goal or threat, then the groups' efficacy is not limited by the variance of the group. This supports research from Charness et al. (2007), since I too demonstrate that aligning individuals through repeated interaction and common purposes increases group-wide cooperation. However, perhaps this null result is observed because high stakes crowd out any incentive for an individual to exhibit any form of biases to others and since altruism has been shown to wane as stakes increase (Ra-

bin, 1993). Moreover, this null result is subject to the identifying factors that I chose to measure and their variation within the sample, and therefore should not be considered a resounding conclusion of no effect. Finally, one should be cautious that this null result may instead be driven by power issues. The standard errors of estimated coefficients to 'zeta' interactions in model three are not too large but one may be hesitant to reject this reason given the small variance in the 'zeta' composite variable.

The main conclusion of this thesis: that individuals may have unstable cooperative phenotypes co-determined by social and private incentives, should not be interpreted as a suggestion that anyone can be a conditional cooperator under the right incentives. This question is beyond the scope of my research. Since the results are an average over all subjects, it is perfectly plausible that some of the observed individuals exhibit constant phenotypes.

In addition to the aforementioned limitations of this study, the external validity of my results is restricted by the strength of my exclusion criteria and identifying assumptions. In particular, I am most concerned with the assumption of strictly exogenous external threats. By excluding individuals who 'flip' alliances, I assume that the external threat is exogenous to those subjects who I include in my panel data set. However, I am only able to exclude individuals who flip during the periods of observation and cannot identify individuals who 'flip' alliances in a period after I finish observing and thus act in anticipation of this eventual flip. If this effect is found then it would jeopardise the results in that my coefficient estimators would be biased since individuals would behave in considering future interactions outside of the defined neighbourhood. Further research should invest resources into strictly prohibiting the flipping of alliances by containing all interactions within the recorded periods.

The source of the threat that one wishes to overcome in this study is not an inanimate concept[18]: it is instead another rival group of individuals. On the one hand, this intergroup competition may create a strong incentive for intragroup coordination in face of the stochastic nature of a human opposition (Bornstein et al., 2002). Conversely, the effect of 'ganging up' guilt may be felt here by the most altruistic individuals. This effect is such that individuals who cooperate may feel guilty upon condemning an individual to a rather large paper loss in potential winnings; this feeling may translate to a utility loss. Thus those who are the most pro-social may be faced with an internal conflict between eliciting a cooperative equilibrium within their neighbourhood but not seeming too harsh on another individual. That being said, I am not too concerned of this effect since it is ex ante prescribed by contestants that they must vote for another individual and condemn them to eviction at some point during the game. Instead, if one does not abide by this voting behaviour, then one is most prone to suffering the loss oneself. Further, it is concluded by Rabin (1993) that altruism is exhibited less frequently the greater the stake of the interaction. Hence, since I am observing interactions under very high stakes with essential voting, I am little concerned with guilt effects, however the effect may be muted when intergroup competition is not present and the rival group is replaced with an environmental threat. Accordingly, it would be fruitful for future research to replicate this result using non-group competition.

Further, the internal validity of my results is restricted by the robustness of my definition of a neighbourhood. On the one hand, a neighbourhood, by definition is self-contained and closed, however, I observe many instances of inter-neighbourhood interaction. In order for my results to show that group-wide cooperation strategies are dependent on social norms and private incentives, I need to be able to define a group within which this decision occurs. A weak definition of the neighbourhood

[18]For example an institutional threat to humanity such as climate change or inequality.

may underestimate the average observed sensitivity to the social norm since either individuals are less concerned with the decisions of those in their circle or are even intimidated by the interactions of these individuals. Further research should aim to isolate this neighbourhood through experimental definitions to ensure strict containment. Finally, it may be the case that neighbourhoods vary in their stability by the presence of sub-neighbourhoods. Thus an individual within a neighbourhood may be more sensitive to the average decision of those in their sub-neighbourhood than others in their sub-neighbourhood. This effect is directly observed by Kok et al. (2020) in Tanzania, such that individuals are closest to their sub-neighbourhood members and so respond to the average decision of this group more reciprocally than the average decision of the neighbourhood.

Moreover, an additional source of concern regarding the internal validity of this thesis is the variability of the external threat within my data set. In fact, while I refer to threats which are not incentive compatible, only 8.6% of my observation reflect such a situation where the value of 'Log Threat' exceeds the incentive compatibility threshold of 0.8625. If one considers the theoretically estimate of the incentive compatibility threshold, 18% of observations represent a relative threat at or above a value of 1. Further research may wish to test the internal validity of my assumptions by increasing the relative representation of incentive incompatible observations.

This thesis contributes to three strands of the literature. First and most importantly, I provide evidence for the instability of individual reciprocity according to environmental factors. I demonstrate that individuals adjust their strategies as cooperation ceases to be an equilibrium strategy (Dreber et al., 2014; Arechar et al., 2018; Embrey et al., 2018). However, an environment in which cooperation is not incentive compatible does not strictly prohibit cooperation but does diminish one's incentive to unconditionally cooperate and increase one's sensitivity to the average action of others which could lead to a shift in strategy towards blind defection (Dreber et al., 2014) or could lead to the adaptation to a new defection norm (Kandori, 1992; Traxler and Spichtig, 2010). Thus I show that cooperation can be achieved through framing the cooperative act as incentive compatible or through demonstrating the presence of a strong social norm. However, these effects are interdependent and the relative importance of each policy approach depends on the external environment.

Second, I contribute to the literature on the dynamics of social norm influence on maintaining equilibria even when they are no longer incentive compatible (Smerdon et al., 2016; Breitmoser, 2015). Thus I emphasise the importance of the power of social norm beliefs on achieving pro-social outcomes and add to the argument that policies which enhance these beliefs through increasing informational transparency for example, may improve group-wide coordination on cooperation (Dal Bó and Fréchette, 2018).

Third, I contribute to the literature on the modelling of n-person games with endogenously determined repeated interactions. This literature is an important study since it mirrors many realistic scenarios involving group interaction under evolving parameters (Duffy and Xie, 2016; Fréchette and Yuksel, 2016). While I do not explicitly measure the effect of group size, I propose a game that incorporates endogenous game exit and social norm sensitivity into coordination decisions.

9 Conclusion

Are individual cooperative phenotypes stable across different environments? This thesis proposes a novel study of this research question using data from the US TV game show *Survivor*. Upon

constructing a series of fixed effect linear probability models which are adapted to the framework of the Experience Weighted Attraction with Norm Psychology Model (Camerer and Ho, 1999; Realpe-Gómez et al., 2019), I am able to identify a continuous behavioural response to evolving social norms and individual incentives to cooperate.

The main results are as follows. First, it is shown that external threats and social norms play a deciding role in the absolute and relative attractiveness of the act of cooperation. Second, rising external threats on average diminish the rate of unconditional and conditional cooperation. Third, this effect is not linear and in fact dependent on whether a certain incentive compatibility constraint is satisfied such that cooperation is a private equilibrium choice according to the theoretical model. When this constraint is satisfied, an increasing external threat translates to a behavioural shift from conditional cooperation towards unconditional cooperation. Under a weak social norm of cooperation, this results in an average increase in the rate of cooperation. However, under a strong social norm of cooperation, this translates to a slight fall in the average rate of cooperation. When this constraint is not satisfied, individuals in fact become more sensitive to the average decision of others in their interacting group, termed the neighbourhood. Thus the results demonstrate that under strong social norms, individuals tend to continue cooperating as the external threat rises. In the case of weak social norms, cooperation is prone to unravel as individuals adjust their beliefs and respond to the average decision which tends towards the defection pole. Finally, I do not find any difference in this effect according to group composition holding group size constant.

This thesis presents a novel case which distinguishes between two commonly merged effects: social incentives and private incentives, and facilitates the identification of factors that are important in achieving high cooperation rates in evolving environments. The results present an argument for the importance of social norm research as a tool in understanding group endogenous behaviour. Further, the study proposes that reciprocity itself can be measured on a continuous scale using a probability model.

While the study using game show data is not nuanced, the subjectivity of the data set renders several precautions. The internal validity of such a study is questionable since the data does not strictly define interactions nor objective groups. Further external validity hinges on the composition of the panel data set regarding the assumed exogeneity of the external threat under the condition that individuals do not choose to mould their exposure to risks.

Further research should be directed to replicating the dynamic transition of conditional cooperation strategies under evolving environmental parameters. On top of the avenues which have been hinted at, replication studies should focus on achieving truly randomly selected subjects faced with inanimate threats to confirm the internal validity of this thesis. Research may be interested in deriving an exogenous definition of this risk amount to systematically remove risk from the decision. Finally, research may be drawn into the surprising null result of group composition and be interested in conducting a follow up study which permits the size of groups to evolve permitting the identification of group size on the instability of conditional cooperation strategies.

References

Ackermann, K. and Murphy, R. (2019), 'Explaining cooperative behavior in public goods games: How preferences and beliefs affect contribution levels', *Games* **10**(1), 15.

Akerlof, G. A. and Kranton, R. E. (2000), 'Economics and identity*', *Quarterly Journal of Economics* **115**(3), 715–753.

Ali, S. N. and Miller, D. A. (2016), 'Ostracism and forgiveness', *American Economic Review* **106**(8), 2329–2348.

Arechar, A., Kouchaki, M. and Rand, D. (2018), 'Examining spillovers between long and short repeated prisoner's dilemma games played in the laboratory', *Games* **9**(1), 5.

Axelrod, R. and Hamilton, W. D. (1981), 'The evolution of cooperation', *Science* **211**(4489), 1390–1396.

Balikci, A. (1970), *The Netsilik Eskimo*, Natural History Press.

Balliet, D. (2009), 'Communication and cooperation in social dilemmas: A meta-analytic review', *Journal of Conflict Resolution* **54**(1), 39–57.

Balliet, D., Li, N. P., Macfarlan, S. J. and Vugt, M. V. (2011), 'Sex differences in cooperation: A meta-analytic review of social dilemmas.', *Psychological Bulletin* **137**(6), 881–909.

Balliet, D., Tybur, J. M., Wu, J., Antonellis, C. and Lange, P. A. M. V. (2016), 'Political ideology, trust, and cooperation', *Journal of Conflict Resolution* **62**(4), 797–818.

Bednarik, R. D. (2008), 'The domestication of humans', *anthropologie* **46**(1), 1–17.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Zandt, T. V., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. and Johnson, V. E. (2018), 'Redefine statistical significance', *Nature Human Behavior* **2**, 6–10.

Bernhard, H., Fehr, E. and Fischbacher, U. (2006), 'Group affiliation and altruistic norm enforcement', *American Economic Review* **96**(2), 217–221.

Bernheim, B. D. (1994), 'A theory of conformity', *The Journal of political economy* **102**(5), 841–877.

Bicchieri, C. (1990), 'Norms of cooperation', *Ethics* **100**(4), 838–861.

Bigoni, M., Bortolotti, S., Casari, M., Gambetta, D. and Pancotto, F. (2016), 'Amoral familism, social capital, or trust? the behavioural foundations of the italian north-south divide', *The Economic journal (London)* **126**(594), 1318–1341.

Billingsley, J., Gomes, C. M. and Mccullough, M. E. (2018), 'Implicit and explicit influences of religious cognition on dictator game transfers', *Royal Society Open Science* **5**(8), 170238.

Blonski, M., Ockenfels, P. and Spagnolo, G. (2011), 'Equilibrium selection in the repeated prisoners dilemma: Axiomatic approach and experimental evidence', *American Economic Journal: Microeconomics* **3**(3), 164–192.

Blurred, R. (2010), 'Survivor contestant contract: the waivers and agreements that cast members and families sign'.

Bornstein, G., Gneezy, U. and Nagel, R. (2002), 'The effect of intergroup competition on group coordination: an experimental study', *Games and Economic Behavior* **41**(1), 1–25.

Bowles, S. and Gintis, H. (2011), *A Cooperative Species*, Princeton University Press, Princeton and NJ and USA.

Brandts, J. and Charness, G. (2011), 'The strategy versus the direct-response method: a first survey of experimental comparisons', *Experimental Economics* **14**(3), 375–398.

Breitmoser, Y. (2015), 'Cooperation, but no reciprocity: Individual strategies in the repeated prisoner's dilemma', *American Economic Review* **105**(9), 2882–2910.

Brock, W. A. and Durlauf, S. N. (2001), 'Discrete choice with social interactions', *The Review of economic studies* **68**(2), 235–260.

Camerer, C. and Ho, T. H. (1999), 'Experience-weighted attraction learning in normal form games', *Econometrica* **67**(4), 827–874.

Cameron, A. C. and Miller, D. L. (2015), 'A practitioner's guide to cluster-robust inference', *The Journal of human resources* **50**(2), 317–372.

Charness, G. and Chen, Y. (2020), 'Social identity, group behavior, and teams', *Annual Review of Economics* **12**(1), 691–713.

Charness, G., Cobo-Reyes, R. and Jiménez, N. (2014), 'Identities, selection, and contributions in a public-goods game', *Games and Economic Behavior* **87**, 322–338.

Charness, G., Rigotti, L. and Rustichini, A. (2007), 'Individual behavior and group membership', *American Economic Review* **97**(4), 1340–1352.

Chen, Y., Li, S. X., Liu, T. X. and Shih, M. (2014), 'Which hat to wear? impact of natural identities on coordination and cooperation', *Games and Economic Behavior* **84**, 58–86.

Cooper, D. J. and Kühn, K.-U. (2016), 'Communication and cooperation: A methodological study', *Southern economic journal* **82**(4), 1167–1185.

Dal Bó, P. (2005), 'Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games', *American Economic Review* **95**(5), 1591–1604.

Dal Bó, P. and Fréchette, G. R. (2011), 'The evolution of cooperation in infinitely repeated games: Experimental evidence', *The American economic review* **101**(1), 411–429.

Dal Bó, P. and Fréchette, G. R. (2018), 'On the determinants of cooperation in infinitely repeated games: A survey', *Journal of Economic Literature* **56**(1), 60–114.

Dawes, R. M. and Thaler, R. H. (1988), 'Anomalies: Cooperation', *Journal of Economic Perspectives* **2**(3), 187–197.

Deb, J. and González-Díaz, J. (2019), 'Enforcing social norms: Trust-building and community enforcement', *Theoretical Economics* **14**(4), 1387–1433.

Dreber, A., Fudenberg, D. and Rand, D. G. (2014), 'Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics', *Journal of economic behavior organization* **98**, 41–55.

Drouvelis, M., Malaeb, B., Vlassopoulos, M. and Wahba, J. (2021), 'Cooperation in a fragmented society: Experimental evidence on syrian refugees and natives in lebanon', *Journal of economic behavior organization* **187**, 176–191.

Drukker, D. (2003), 'Testing for serial correlation in linear panel-data models', *The STATA journal* **2**, 168–177.

Duffy, J. and Xie, H. (2016), 'Group size and cooperation among strangers', *Journal of Economic Behavior Organization* **126**, 55–74.

Eagly, A. H. and Mladinic, A. (1994), 'Are people prejudiced against women? some answers from research on attitudes, gender stereotypes, and judgments of competence', *European Review of Social Psychology* **5**(1), 1–35.

Embrey, M., Fréchette, G. R. and Yuksel, S. (2018), 'Cooperation in the finitely repeated prisoner's dilemma', *The Quarterly journal of economics* **133**(1), 509–551.

Erikson, R. and Goldthorpe, J. (1992), *The constant flux : a study of class mobility in industrial societies*, 1 edn, Clarendon Press, Oxford [England].

EU (2016), European Union.
URL: *https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679*

Fehr, E. and Gächter, S. (2000*a*), 'Cooperation and punishment in public goods experiments', *American Economic Review* **90**(4), 980–994.

Fehr, E. and Gächter, S. (2000*b*), 'Fairness and retaliation', *The Economics of Reciprocity, Giving and Altruism* p. 153–173.

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993), 'Does fairness prevent market clearing? an experimental investigation', *The Quarterly Journal of Economics* **108**(2), 437–459.

Fehr, E. and List, J. A. (2003), 'The hidden costs and returns of incentives - trust and trustworthiness among ceos', *SSRN Electronic Journal* .

Fehr, E. and Schmidt, K. M. (1999), 'A theory of fairness, competition, and cooperation', *The Quarterly journal of economics* **114**(3), 817–868.

Fischbacher, U., Gächter, S. and Fehr, E. (2001), 'Are people conditionally cooperative? evidence from a public goods experiment', *Economics letters* **71**(3), 397–404.

Fréchette, G. R. and Yuksel, S. (2016), 'Infinitely repeated games in the laboratory: four perspectives on discounting and random termination', *Experimental Economics* **20**(2), 279–308.

Gilovich, T., Griffin, D. W. and Kahneman, D. (2009), *Heuristics and biases: the psychology of intuitive judgement*, Cambridge University Press.

Gneezy, U., Leibbrandt, A. and List, J. A. (2016), 'Ode to the sea: Workplace organizations and norms of cooperation', *The Economic journal (London)* **126**(595), 1856–1883.

Haag, M. and Lagunoff, R. (2007), 'On the size and structure of group cooperation', *Journal of Economic Theory* **135**(1), 68–89.

Helénsdotter, R. (2019), Experimental Evidence on Cooperation and Political Affiliation and and Group Size, PhD thesis.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q. and Tracer, D. (2005), '"economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies', *The Behavioral and brain sciences* **28**(6), 795–815.

Henrich, J. P. (2016), *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*, Princeton University Press.

Henrich, N. and Henrich, J. P. (2007), *Why humans cooperate: a cultural and evolutionary explanation*, Oxford University Press.

Hill, T. D., Davis, A. P., Roos, J. M. and French, M. T. (2020), 'Limitations of fixed-effects models for panel data', *Sociological perspectives* **63**(3), 357–369.

Hoffman, E., Mccabe, K. A. and Smith, V. L. (2000), 'Social distance and other-regarding behavior in dictator games', *Bargaining and Market Behavior* p. 127–138.

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied logistic regression, third edition*, 3rd ed. edn, John Wiley and Sons, Hoboken and NJ.

Irlenbusch, B., Rilke, R. M. and Walkowitz, G. (2018), 'Designing feedback in voluntary contribution games: the role of transparency', *Experimental Economics* **22**(2), 552–576.

Kahneman, D. (2013), *Thinking, fast and slow*, Farrar, Straus and Giroux.

Kahneman, D., Knetsch, J. L. and Thaler, R. H. (1991), 'Anomalies: The endowment effect, loss aversion, and status quo bias', *Journal of Economic Perspectives* **5**(1), 193–206.

Kandori, M. (1992), 'Social norms and community enforcement', *The Review of Economic Studies* **59**(1), 63.

Karlan, D. (2017), 'Survivor: Three principles of economics lessons as taught by a reality television show', *The Journal of economic education* **48**(3), 224–228.

Kok, L., Oosterbaan, V., Stoker, H. and Vyrastekova, J. (2020), 'In-group favouritism and social norms: Public goods experiments in tanzania', *Journal of Behavioral and Experimental Economics* **85**, 101509.

Kumar, M. M., Tsoi, L., Lee, M. S., Cone, J. and Mcauliffe, K. (2021), 'Nationality dominates gender in decision-making in the dictator and prisoner's dilemma games', *Plos One* **16**(1).

Kurokawa, S., Wakano, J. Y. and Ihara, Y. (2018), 'Evolution of groupwise cooperation: Generosity, paradoxical behavior, and non-linear payoff functions', *Games* **9**(4), 100.

Lee, R. B. (1979), *The !Kung San: Men and Women and Work in a Foraging Society*, Cambridge University Press.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical analysis with missing data*, Wiley.

Martin, C. F., Bhui, R., Bossaerts, P., Matsuzawa, T. and Camerer, C. (2014), 'Chimpanzee choice rates in competitive games match equilibrium game theory predictions', *Scientific Reports* **4**(1).

Martinangeli, A. F. M. and Martinsson, P. (2020), 'We and the rich: Inequality, identity and cooperation', *Journal of economic behavior organization* **178**, 249–266.

Marwell, G. and Ames, R. E. (1981), 'Economists free ride, does anyone else?', *Journal of Public Economics* **15**(3), 295–310.

Mccullough, M. E., Swartwout, P., Shaver, J. H., Carter, E. C. and Sosis, R. (2016), 'Christian religious badges instill trust in christian and non-christian perceivers.', *Psychology of Religion and Spirituality* **8**(2), 149–163.

Mengel, F. (2018), 'Risk and temptation: A meta-study on prisoner's dilemma games', *The Economic journal (London)* **128**(616), 3182–3209.

Murnighan, J. K. and Roth, A. E. (1983), 'Expecting continued play in prisoners dilemma games', *Journal of Conflict Resolution* **27**(2), 279–300.

Nagatsu, M., Larsen, K., Karabegovic, M., Szekely, M., Monster, D. and Michael, J. (2018), 'Making good cider out of bad apples -signaling expectations boosts cooperation among would-be free riders', *Judgment and decision making* **13**(1), 137–149.

Nikiforakis, N. and Normann, H.-T. (2007), 'A comparative statics analysis of punishment in public-good experiments', *Experimental Economics* **11**(4), 358–369.

Norenzayan, A. and Shariff, A. F. (2008), 'The origin and evolution of religious prosociality', *Science* **322**(5898), 58–62.

Nowak, M. A. and Sigmund, K. (1992), 'Tit for tat in heterogeneous populations', *Nature* **355**(6357), 250–253.

Olson, M. (1971), *The logic of collective action*, Vol. 124, 2. print. edn, Harvard Univ. Press, Cambridge and Mass. [u.a.].

Parrotta, P., Pozzoli, D. and Pytlikova, M. (2014), 'The nexus between labor diversity and firm's innovation', *Journal of population economics* **27**(2), 303–364.

Pillutla, M. M. and Chen, X.-P. (1999), 'Social norms and cooperation in social dilemmas: The effects of context and feedback', *Organizational behavior and human decision processes* **78**(2), 81–103.

Post, T., van den Assem, J, M., Baltussen, G. and Thaler, R. H. (2008), 'Deal or no deal? decision making under risk in a large-payoff game show', *The American economic review* **98**(1), 38–71.

Potoms, T. and Truyts, T. (2016), 'On symbols and cooperation', *SSRN Electronic Journal* .

Proto, E., Rustichini, A. and Sofianos, A. (2019), 'Intelligence, personality, and gains from cooperation in repeated interactions', *The Journal of political economy* **127**(3), 1351–1390.

Rabin, M. (1993), 'Incorporating fairness into game theory and economics', *The American economic review* **83**(5), 1281–1302.

Realpe-Gómez, J., Vilone, D., Andrighetto, G., Nardin, L. and Montoya, J. (2019), 'Learning dynamics and norm psychology supports human cooperation in a large-scale prisoner's dilemma on networks'.

Reuben, E. and Suetens, S. (2011), 'Revisiting strategic versus non-strategic cooperation', *Experimental Economics* **15**(1), 24–43.

Romero, J. and Rosokha, Y. (2018), 'Constructing strategies in the indefinitely repeated prisoner's dilemma game', *European economic review* **104**, 185–219.

Rossetti, C. S. L., Hilbe, C. and Hauser, O. P. (forthcoming), '(mis)perceiving cooperativeness', *Current opinion in psychology* **43**, 151–155.

Roth, A. E. and Murnighan, J. (1978), 'Equilibrium behavior and repeated play of the prisoners dilemma', *Journal of Mathematical Psychology* **17**(2), 189–198.

Rustagi, D., Engel, S. and Kosfeld, M. (2010), 'Conditional cooperation and costly monitoring explain success in forest commons management', *Science (American Association for the Advancement of Science)* **330**(6006), 961–965.

Schaffer, M. and Stillman, S. (2010), 'xtoverid: Stata module to calculate tests of overidentifying restrictions after xtreg and xtivreg and xtivreg2 and xthtaylor '.

Smerdon, D., Offerman, T. and Gneezy, U. (2016), 'Everybody's doing it: On the emergence and persistence of bad social norms', *Tinbergen Institute, Amsterdam and Rotterdam* .

Smith and Adam (1759), 'The theory of moral sentiments', *The Glasgow Edition of the Works and Correspondence of Adam Smith and Vol. 1: The Theory of Moral Sentiments* .

Sobel, J. (2005), 'Interdependent preferences and reciprocity', *Journal of Economic Literature* **43**(2), 392–436.

Tajfel, H., Billig, M. G., Bundy, R. P. and Flament, C. (1971), 'Social categorization and intergroup behaviour', *European journal of social psychology* **1**(2), 149–178.

Thöni, C. and Volk, S. (2018), 'Conditional cooperation: Review and refinement', *Economics Letters* **171**, 37–40.

Tognetti, A., Berticat, C., Raymond, M. and Faurie, C. (2012), 'Sexual selection of human cooperative behaviour: An experimental study in rural senegal', *PLoS ONE* **7**(9).

Tognetti, A., Dubois, D., Faurie, C. and Willinger, M. (2016), 'Men increase contributions to a public good when under sexual competition', *Scientific reports* **6**(1), 29819.

Traxler, C. and Spichtig, M. (2010), 'Social norms and the indirect evolution of conditional cooperation', *Journal of Economics* **102**(3), 237–262.

United States Census Bureau (2021), 'Census regions and divisions of the united states'.
URL: *https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf*

Wooldridge, J. M. (2002), *Econometric analysis of cross section and panel data*, MIT Press, Cambridge and Mass. [u.a.].

Wooldridge, J. M. (2013), *Introductory econometrics*, 5. ed. and internat. ed. edn, South-Western Cengage Learning, Mason and Ohio u.a.

Yu, T., Chen, S.-H. and Li, H. (2015), 'Social norms, costly punishment and the evolution of cooperation', *Journal of Economic Interaction and Coordination* **11**(2), 313–343.

10 Appendix 1: Proof of equation (3)

Begin with equation (2) that defines the equilibrium average choice of players in individual i's neighbourhood. First define k as the number of cooperators in equilibrium such that $\bar{m}_i^*$ can be defined as the following:

$$\bar{m}_i^* = \frac{2k - N + 1}{N - 1}$$

Allow k to be a binomial variable with parameters $n = N - 1$ and $p = F(2\lambda_i J\bar{m}_i^* - c + v(1))$ such that the expected number of cooperators is given as $E(k) = np = (N-1)F(2\lambda_i J\bar{m}_i^* - c + v(1))$. Substituting $E(k)$ into equation (4), the calculation is now:

$$\bar{m}_i^* = \frac{(N-1)(p-1)}{N-1} = (p-1) = F(2\lambda_i J\bar{m}_i^* - c + v(1)) - 1$$