Preferences and Cooperation

Alexander Harris

St. John's College

***

This dissertation is submitted to

the University of Cambridge

for the degree of

Doctor of Philosophy

***

Supervisor: Professor Robert Evans

***

Faculty of Economics

Preferences and Cooperation

Alexander Harris

Abstract

Chapter 1: Evolution of reciprocator preferences when agents can pay for information

A benchmark result in the evolutionary games literature is that a preference for reciprocity will evolve if preferences are observable (at zero cost), since reciprocators can cooperate with each other rather than with materialists, thereby achieving a fitness advantage. I investigate how a preference for reciprocity evolves if individuals can observe an opponent's preferences only by bearing a fitness cost. My main result applies when observing an opponent's type is cheap, but cooperating only gives a modest fitness advantage or the preference for reciprocity is intense. In this case, a preference for reciprocity *cannot* evolve from a small starting share in the mix of preferences, even if discovering an opponent's preferences is arbitrarily cheap. This is in sharp contrast to the benchmark result.

Chapter 2: A theory of conditional cooperation on networks (with Julien Gagnon)

Chapter 2 is a study of reciprocity on social networks. We model a group of connected agents who play a one-shot public good game. Some players are materialists and others are reciprocators. We characterise the maximal Nash equilibrium (ME) of this game for any network and a broad class of reciprocal preferences. At the ME, a novel concept, the q-linked set, fully determines the set of players who contribute. We show that influential players are those connected to players who are sufficiently interconnected, but not too much. Finally, we study the decision of a planner faced with an uncertain type profile who designs the network to

maximise expected contributions. The *ex ante* optimal network comprises isolated cliques of degree $k^* \geq 1$, with k^* decreasing with the incidence of materialists. We discuss an important application of our results: the workplace.

Chapter 3: Ideological games

Chapter 3 is a theory of ideology. I define a preference type to be a set of first-order preferences over the outcomes of a 'game of life' Γ, together with a set of ('meta-') preferences over all players' first-order preferences. Players can influence each other's preferences via costly investment: if player A invests and B does not, B's preferences becomes those of A. Players may invest for instrumental reasons (i.e. to achieve better outcomes in Γ) or 'ideological' reasons (i.e. they want their opponents to have the same preferences they do). I characterise 'strongly ideological', 'weakly ideological' and 'pragmatic' types. Weakly ideological types wish to preserve their own type, as do strongly ideological types, who also seek to convert others. A pragmatic player, in contrast, is willing to have her type changed if her new type would prefer the resulting equilibrium of Γ to the status quo. I show that if two players of different ideological types meet, there is an equilibrium investment profile with lower aggregate welfare than the no-invest profile. If at least one type is strongly ideological, there is a unique such equilibrium. Finally, a 'perfectly ideological' type is a strongly ideological type which, if held by all players, results in the best outcome of Γ as judged by that type. If a perfectly ideological player plays a pragmatic player, aggregate welfare is always greater than in the no-invest profile.

Preferences and Cooperation

Alexander Harris

Contents

Preface

I hereby declare that dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. Collaborative work is limited to Chapter 2 (co-authored with Julien Gagnon, University of Cambridge). I further declare that this dissertation is not substantially the same as any that I have submitted or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other university, and that this dissertation does not exceed the prescribed word limit of 60,000 words.

Alexander Harris

November 2018

Acknowledgements

I would like to thank the following people for their help and support.

I have been especially lucky to have had such a patient, insightful and supportive supervisor in Robert Evans, to whom I would like to express my deep gratitude. He has been continually generous with his time in guiding my intellectual development and spending time in his company has been a true pleasure. I am also most grateful to my research advisor, Aytek Erdil, and to many current and former faculty members at Cambridge including in particular Matt Elliot, Eduardo Gallo, Sanjeev Goyal, Charles Roddie, Hamid Sabourian and Oliver Walker, for their thoughtful advice, constructive criticism and kind encouragement.

I would also like to thank my fellow doctoral students for making my working and social environments so pleasant. Hugely important to me during my time at Cambridge was my friendship with Julien Gagnon, Sulaiman Ijaz and Alex Ross. Additionally, I would like to Julien for being a fantastic coauthor. A number of my other friends have also been most supportive during my studies, including my very good friend James Millen.

I am very grateful for the considerable financial support I received from the Economic and Social Research Council for my graduate studies.

Finally, I am forever grateful to my wonderful family, especially my parents Avril and Nick and my wife Emma, who have been a constant source of much love and support. I thank Emma in particular for doing everything she could in recent years to help me pursue my goal of doing postgraduate research.

Introduction and summary

Two related and overarching topics of enquiry in my thesis are how non-materialistic preferences support cooperation and how such preferences originate. There is ample experimental evidence that some people (*reciprocators*) have a preference for mutual cooperation, whereas other players (*materialists*) prefer to free-ride in such situations. I study the evolutionary origins of a preference for reciprocity in Chapter 1. In Chapter 2, the focus is cooperation on networks. Reciprocators are replaced with an analogous type of player – *conditional cooperators* – who are influenced towards cooperation and free-riding, respectively, by neighbours who contribute and free-ride in a public good game. Finally, in Chapter 3, I offer a theory of ideology in which individuals can seek to change each other's preferences. A more detailed chapter-by-chapter summary is as follows.

Chapter 1: Evolution of reciprocator preferences when agents can pay for information

A benchmark result in the evolutionary games literature is that a preference for reciprocity will evolve if preferences are observable (at zero cost), since reciprocators can cooperate with each other rather than with materialists, thereby achieving a fitness advantage. I investigate how a preference for reciprocity evolves if individuals can observe an opponent's preferences only by bearing a fitness cost. My main result applies when observing an opponent's type is cheap, but cooperating only gives a modest fitness advantage or the preference for reciprocity is intense. In this case, a preference for reciprocity *cannot* evolve from a small starting share in the mix of preferences, even if discovering an opponent's preferences is arbitrarily cheap. This is in sharp contrast to the benchmark result. I offer an interpretation of paying to observe preferences, using the sociological concept of 'social distance': individuals can adopt a policy of only interacting with others who are close enough socially to them. I also show that individuals

will pay to observe each other's preferences only if the population share of reciprocators takes an intermediate value. Under the 'social distance' interpretation of my model, this latter result generates a prediction: societies with intermediate levels of cooperation should be more fragmented into identity groups than societies with high or low such levels.

Chapter 2: A theory of conditional cooperation on networks (with Julien Gagnon)

Chapter 2 is a study of reciprocity on social networks. We model a group of connected agents who play a one-shot public good game. Some players are materialists and others are reciprocators. We characterise the maximal Nash equilibrium (ME) of this game for any network and a broad class of reciprocal preferences. At the ME, a novel concept, the q-linked set, fully determines the set of players who contribute. We provide a characterisation of players' influence at equilibrium, and show that influential players are those connected to players who are sufficiently interconnected, but not too much. Finally, we study the decision of a planner faced with an uncertain type profile who designs the network to maximise expected contributions. We find that the *ex ante* optimal network comprises isolated cliques of degree $k^* \geq 1$, with k^* decreasing with the incidence of materialists. We discuss evidence in support of our results in the context of one important application: the workplace.

Chapter 3: Ideological games

Chapter 3 is a theory of ideology. I define a preference type to be a set of first-order preferences over the outcomes of a 'game of life' Γ , together with a set of ('meta-') preferences over all players' first-order preferences. Players can influence each other's preferences via costly investment: if player A invests and B does not, B's preferences becomes those of A. Players may invest for instrumental reasons (i.e. to achieve better outcomes in Γ) or 'ideological' reasons (i.e. they want their opponents to have the same preferences they do). I characterise 'strongly ideological', 'weakly ideological' and 'pragmatic' types. Strongly ideological types

always seek to convert others and to preserve their own type, whereas weakly ideological types do not always seek to convert others but do wish to preserve their own type. A pragmatic player, in contrast, is willing to have her type changed if her new type would rank the resulting equilibrium of Γ more highly than she currently ranks the status quo equilibrium. I show that if two players of different ideological types meet, there is an equilibrium investment profile in which aggregate welfare is lower than in the no-invest profile. If at least one type is strongly ideological, there is a unique such equilibrium investment profile. I then introduce a 'perfectly ideological' type, a strongly ideological type which, if held by all players, results in the best outcome of Γ as judged by that type. If a perfectly ideological player plays a pragmatic player, there is a unique equilibrium in which aggregate welfare exceeds that of the no-invest profile. Finally, I sketch how ideological games provide a new mechanism through which a preference for cooperation might evolve, expanding the framework of Chapter 1 to incorporate the notion of cultural or memetic (as opposed to genetic) evolution.

Chapter 1 Evolution of reciprocator preferences when agents can pay for information

Alexander Harris[1]

1.1 Introduction

Human beings cooperate in many situations, even when doing so yields no personal gain in the present or the future. A wealth of experimental evidence establishes that this is so because some people have a preference for reciprocity. In other words, some people are *reciprocators*: they willingly bear material costs to bestow a benefit on others (i.e. cooperate) precisely when they believe that others will do likewise. In contrast, self-interested (*materialist*) players never want to cooperate with others, though they would like others to cooperate with them. Models in evolutionary game theory explain how, and under what conditions, a preference for reciprocity can evolve. If players know each other's preferences, reciprocators can cooperate with each other rather than with materialists, thereby achieving a fitness advantage over the latter. A benchmark result in the literature is that a preference for reciprocity will evolve if preferences are observable with a high enough fixed probability, and if cooperating bestows a sufficiently large benefit on one's opponent. The model I introduce in this chapter reveals a major constraint on this result, however.

I investigate how a preference for reciprocity will evolve if information about each other's preferences is endogenous. To do this, I consider in particular what happens if a "discovery

[1]I am very grateful to Robert Evans for his suggestions and for numerous and lengthy discussions on this topic. I would also like to thank Aytek Erdil, Julien Gagnon, Edoardo Gallo, Hamid Sabourian, seminar participants in Cambridge and conference participants at King's College London for their helpful comments. I am grateful for the financial support I received from the Economic and Social Research Council.

technology" is introduced, meaning individuals can pay (i.e. bear a fitness cost) to improve their chance of discovering an opponent's preferences. It turns out that such a technology may help or hinder the evolution of a preference for reciprocity, depending on parameter values. If gaining information is expensive, then players will not use the technology, so it will make no difference. If gaining information about an opponent's type is cheap, and mutual cooperation has large fitness benefits, then the discovery technology supports the evolution of a preference for reciprocity, as it helps reciprocators discern between different types of opponent.

The main result in this chapter characterises when the discovery technology undermines the evolution of a preference for reciprocity, which happens when information is cheap but the fitness benefit b of mutual cooperation is relatively low or the preference for reciprocity is intense (Theorem 1.1). Even if players can observe each other's preferences perfectly at arbitrarily low cost, a preference for reciprocity cannot evolve from a small starting share in the mix of preferences. The reason is that having a preference for reciprocity makes individuals willing to 'overpay' to use the technology. Their willingness stems from the fact they value mutual cooperation above the direct fitness benefit it gives them, which is the very feature of a preference for reciprocity needed for it to evolve at all. This result is in sharp contrast to the theory of preference evolution under costless observability of preferences, in which a preference for reciprocity evolves for all permissible values of b. In other words, my model yields a significant negative result for the theory that preferences for reciprocity, which drive much cooperative behaviour, evolved via preference observability.

A related (and counterintuitive) result in my model is that if b is relatively low, or the preference for reciprocity is intense, then the evolutionary success of a preference for reciprocity is increasing in the cost of observing others' preferences. Despite the fact that lower costs of information help reciprocators *ceteris paribus*, reciprocators respond by incurring such costs at a lower population share, where they have less of an inherent fitness advantage over materialists. This further limits how far they can grow their share of the population.

A second line of enquiry is: under what general circumstances will individuals use the

discovery technology? The answer is intuitive: when ε, the share of individuals with a preference for reciprocity, takes an intermediate value (Proposition 1.1). However, the reasoning is subtle. The incentive for reciprocators to use the technology depends on what they plan to do if they do not learn their opponent's type. At high population shares ε, reciprocators 'blindly' cooperate when ignorant of their opponent's preferences, since it is likely a player of unknown preferences is in fact a fellow reciprocator. For the same reason, at low population shares, reciprocators do not blindly cooperate. The reason to use the discovery technology is therefore entirely different at low population shares (where the incentive is to detect other reciprocators, and therefore increases in ε) compared with high population shares (where the technology is useful in helping detect materialists, an incentive decreasing in ε). Interestingly, however, despite the very different incentives to use the technology depending on ε, reciprocators only use the technology over a contiguous region of population shares.

I offer two interpretations of the discovery technology. One is that it reflects a wide range of situations where people may expend time and effort to try to discern each other's motives or character. For example, while criminal trials typically seek to find out who has a committed a crime, in many cases they also seek to discern whether the person was 'of good character', to establish a motive or to ascertain whether the crime was committed intentionally. Other examples include credit provision and trade, where information about a potential partner's preferences may establish the trust necessary for a trading relationship. In Rohner, Thoenig and Zilbrotti (2013), for example, traders may acquire such information at a cost. In general, the discovery technology simply relaxes the (strict) assumption that information about other people's preferences is entirely exogenous.

Another interpretation makes use of the sociological concept of *social distance*, i.e. the extent to which two individuals or groups differ in terms of such attributes as language, social class, religious or ethnic background and nationality. Under this interpretation, individuals can adopt a policy of only interacting with others who are close enough socially to them, making it easier to guess their opponent's preferences. The result from the model that the discovery technology is used at a contiguous region of population shares then yields a pre-

diction. Societies with intermediate levels of cooperation should be more fragmented into identity groups – e.g. towns and villages should be stratified by class or split along religious lines – whereas interactions between people in societies with either low or high levels of cooperation should be less conditioned on social distance.[2]

My model takes as its starting point uniform random pairing to play a Prisoner's Dilemma (PD) in fitness payoffs, where all players have a non-zero probability of learning their opponent's type. Different types of reciprocators are possible, as a preference type θ is simply defined by the utility it attaches to mutual cooperation. The higher the type, the greater the utility from mutual cooperation; preferences for reciprocity can thus vary in intensity. For simplicity I restrict attention to the case that each player is either a materialist or is of a particular reciprocator type, and then consider what happens when that particular reciprocator type is allowed to vary. The main novel feature of my model is that the probability with which a player learns her opponent's type takes one of two (non-zero) values. Players can choose to "research" their opponent, i.e. to pay a cost so as to learn the opponent's type with the higher probability rather than the lower probability.[3]

I first examine under what conditions players do research, and when they cooperate. To this end, I select a unique equilibrium, namely the one that maximises cooperation, and consider different type distributions as characterised by the population share of reciprocators. Specifically, I compare the model including the discovery technology with an alternative version in which there is no discovery technology: instead, there is a single, fixed probability of type revelation. I then analyse the model from an evolutionary perspective, via the novel concept of reciprocators' *attainable share*. This is the largest reciprocator population share below which, in the maximal cooperation equilibrium, reciprocators enjoy strictly higher average

[2]The key result of my model, which constrains evolutionary accounts of cooperation via observability of preferences, is orthogonal to this prediction. The latter is premised upon the experimentally-confirmed fact that a preference for reciprocity, whatever its origins, is widespread (but not universal). The prediction of a non-monotonic relationship between social fragmentation and cooperation suggests that empirical studies that estimate linear relationships between such variables (e.g. Alesina, Gennaioli and Lovo, 2017) may benefit from a model specification that allows for non-monotonicity.

[3]In either case, the probability of learning an opponent's type is independent of the chance with which one's own type is revealed.

fitness than materialists. It is an attractive metric because it represents how far the reciprocator population grows under any payoff-monotone fitness dynamics (such as the replicator equation), given an arbitrarily small initial population share.

Related literature

The no-technology version of the model, incorporating an interior fixed probability of type revelation, has not itself been previously studied to my knowledge.[4] However, it is closely related to existing work. For instance, Ockenfels (1993) uses the evolutionary approach to examine reciprocator-materialist type distributions, but where types are perfectly observable. If preference types are perfectly observable, and assuming reciprocators coordinate strategies, reciprocators will enjoy higher fitness than materialists at all population shares. This result is a special case of the results of Dekel, Ely and Yilankaya (2007), whose set of preference types comprises all orderings over the outcomes of a generic one-shot two-player symmetric game.

In the rest of this section I outline different attempts in the theoretical literature to account for cooperative behaviour. As I explain below, theories of preference evolution are a valuable way to understand its origins. Additionally, they can explain the origin of moral action.

1.1.1 Theories of cooperative behaviour

There are a number of explanations for the existence of cooperative behaviour between unrelated individuals. These explanations fall into three broad categories: theories involving bounded rationality, theories involving punishment (often in the context of repeated interaction) and theories involving non-material preferences.

The adoption of strategies under bounded rationality can sustain cooperation. Alexander (2007) provides a number of examples of how cooperation may be maintained on networks when players update their strategies according to learning rules. Nowak and May (1992)

[4]While the no-technology results are therefore novel, they are rather narrow in application, as I only consider a two-type distribution. The reason for doing so is to keep the full model, whose primary purpose is to investigate the strategic and evolutionary impact of allowing players to pay for information, tractable.

dispense with strategic play and learning altogether, studying the evolution of cooperation on a lattice in which each individual plays a PD against its neighbours. The relative payoffs determine whether cooperation or defection proliferates, or whether clusters of cooperation and defection emerge side by side. In general, the assumption that individuals have bounded rationality can apply in cases of repeated interaction or one-off interaction. It seems more plausible the more complex the environment.

A second kind of theory allows for agents to punish each other. Punishment strategies have been much-studied in the context of repeated games. In a repeated game setting, it is well known that players can enforce cooperation by credibly threatening to punish unilateral defectors, if players are sufficiently patient. Seminal work in this area of theory includes Axelrod and Hamilton (1981) and Fudenberg and Maskin (1986). In a one-off setting, punishment may or may not be a viable option for players, but experimental evidence shows that even in the absence of punishment options, one-shot cooperation does take place to some extent.[5] In light of this, explaining the presence of cooperative behaviour by appealing to preferences – the third approach listed above – may be desirable for two reasons. Firstly, if agents are acting rationally when they cooperate in a one-shot setting, they must prefer a mutually cooperative outcome over the higher material returns that would accrue if they were to defect. Secondly, explanations in terms of preferences may shed light on how we conceive of people acting morally, construed in a strong sense, i.e. whereby a person internalises a moral rule, rather than simply pays lip-service to it for the sake of expediency.[6]

A third explanation for cooperative behaviour, then, is that agents have preferences that do not simply track individual material payoffs. These "non-materialistic" preferences could, for example, contain a degree of altruism (see e.g. Becker, 1976), describe inequity aversion

[5]There is some debate as to precisely what extent one-off cooperation between strangers who are unable to punish each other is observed experimentally. For a flavour of the debate on this topic, see for example Binmore and Shaked (2010). For a extended review, see Fehr and Falk (2002).

[6]Alexander (2007) sketches how theories in which cooperation arises via learning on networks might explain the existence of moral rules, which function as heuristic devices. People grow accustomed to such rules and may then apply them in one-off settings in which it would be rational for them to act otherwise. He notes that this explanation cannot account for moral intuition, or for moral action in a strong sense, however.

(Fehr and Schmidt, 1999), characterise a desire to reward generosity and punish meanness (Sobel, 2005) or may reflect deliberate adherence to some moral rule.

1.1.2 Evolution of preferences

Much experimental work has been done to explain the proximate causes of observed behaviour by appealing to different types of preferences. To take just one example, Charness and Rabin (2002) use laboratory data to calibrate a utility function that at once represents preferences over the levels and distribution of material payoffs and over other players' adherence to fairness or reciprocity norms. Evolutionary models offer deeper explanations of observed behaviour and (relatedly) the basis of people's moral intuitions. The established approach to quantitative theoretical study of preference evolution, which I take as my starting point, is the *indirect evolutionary approach* of Güth and Yaari (1992).[7] Under this approach, the distribution of preference types in a population leads to equilibrium behaviour, which determines fitness payoffs that in turn determine the evolution of preferences among the population. In the case of perfectly observable preference types, the authors show that preferences that are not aligned with fitness payoffs can be evolutionarily stable. This stability can arise because such preferences enable credible commitment to outcomes that are efficient in fitness payoffs, despite their underlying actions being fitness-dominated.[8]

Güth and Yaari (1992) consider non-materialistic preference types relating to fairness in a bargaining game. Other early papers to use the indirect evolutionary approach examine other specific games, the choice of which is often motivated by the aim of modelling a particular flavour of preference. Bowles and Gintis (1998) examine preferences for strong reciprocity, i.e. punishing free-riding, in a group setting. Sethi and Somnathan (2001) consider interdependent

[7] An alternative evolutionary account of non-materialistic preferences could in fact be predicated on the existence of punishment options. For example, if punishment arose in early societies as a repeated game effect, then over time, people may have internalised norms of cooperation and punishment such that they develop non-materialistic preferences. The internalisation process would be evolutionary in the sense that it would arise from patterns of behaviour in early societies and psychological effects that themselves have an evolutionary basis.

[8] In a similar spirit, action profiles that are not Nash equilibria in fitness payoffs can be sustainable.

preferences. This chapter of my thesis, in studying the effect of endogenous information on the evolution of reciprocator preferences, extends this area of the literature. To my knowledge, the only other investigations of the effect of allowing for costly type discovery relevant to the indirect evolutionary framework are by Güth (1994) and Guttman (2000). Both study sequential games. Güth (1994) allows the first player in a trust game access to a costly, perfect type-detection technology, with types otherwise private information. Guttman (2000) models a market transaction in which players can, under different treatments, variously monitor each other's types and moves, which result in either an all-reciprocator or all-materialist population. I study a simultaneous-move equilibrium in a PD in fitness payoffs, and allow for interior probabilities of type revelation, which allows for a range of interior solutions with respect to population shares.

Ok and Vega-Redondo (2001) consider a general setting in which agents are randomly assigned into a large number of subgroups of the population. The members of the subgroup play a game among themselves. If preferences are not observable, and players can only condition strategies on the distribution of types in the population as a whole, then materialistic preferences, and only materialistic preferences, are stable in a wide class of environments. The intuition for this result is that if preferences are not observable, they cannot facilitate credible commitment, and so agents' fitness is maximised when action profiles are Nash equilibria in fitness, or "Nash outcomes" for short. The authors conclude that this stability condition provides a rationale for the evolution of materialistic preferences, though they note that it covers only the case of random matching and entirely unobservable preferences.

Dekel, Ely and Yilankaya (2007) consider the indirect evolutionary approach in another general environment, in which players are paired to play a symmetric one-shot game, and the set of preference types consist in all orderings over the game's outcomes. The authors investigate the robustness of stability conditions with respect to zero, full and partial observability of preferences. One result is that for all non-zero levels of observability, efficiency in fitness payoffs is necessary for an outcome to be stable. Consequently, introducing even a

small chance of observing preferences can destabilise a Nash outcome if it is not efficient. It follows that if agents are randomly matched to play a PD in fitness payoffs and there is some observability of preferences, it cannot be stable for all agents to defect all the time.[9]

Under uniform random matching, as in my model, types must be at least partly observable for non-materialist preferences to develop among a population. If types are entirely unobservable, the alternative way non-materialist preferences can develop is via a non-uniform matching process.[10] Alger and Weibull (2013), building on the approach of Bergstrom (2003), study how non-materialist preferences can evolve under exogenous assortative matching. Assortative matching means that players of a given preference type are more likely to be matched with a same-type than an other-type opponent. The authors show that when preference types are unobservable, the stable preference type is one whose utility function is a weighting between actual material (fitness) returns and the material returns that would accrue were all other players to adopt the same strategy as the player in question, with the weighting equal to the degree to which the matching process is assortative. The latter component can be thought of as relating to Kant's categorical imperative to act such that one could rationally will that the action were adopted as a general maxim. Assortative matching in the indirect evolutionary framework can therefore account for moral (specifically Kantian) non-materialist preferences.

My model does not prescribe whether the evolutionary selection at work is genetic or cultural. Cultural selection may be less familiar to most readers than natural selection at a

[9] I do not study stability, but rather specify a weaker concept of "attainability", to focus simply on the salient fitness effects in my model. However, the intuition from Dekel, Ely and Yilankaya's result applies in my model, in that for any non-zero baseline probability of type revelation, there is always a high enough reciprocator type who cooperates in some equilibrium at arbitrarily small population shares. In both my model and theirs, no types can do better in fitness terms than the lowest reciprocator type for whom such cooperation is feasible. However, which type is the lowest such type depends on other parameters in the model, which are themselves likely to vary between real-world environments. This variability partly motivates my focus on a broad set of reciprocator types, rather than any specific such type.

[10] One tempting conclusion from the main result of this chapter, which restricts the conditions under which reciprocity can evolve through observability of preferences, is that the alternative evolutionary theory in terms of a matching process can thereby be regarded as epistemically strengthened. However, such an inference is not entirely clear. After all, it seems possible that endogenising the matching process in the latter theory via (even small) frictional costs, as I have in the case of observable types, could impose analogous constraints on type evolution. Indeed, the interpretation I offer of the type discovery technology as representing 'screening' by social distance has the flavour of a matching process. This may be a fruitful area for further research.

genetic level, and it is not obvious what determines "fitness" in a cultural context. A simple way to make sense of this notion is to suppose that fitness payoffs are equivalent to material success, which correlates with the ability of people of a certain preference type to persuade others to adopt their ways, an idea I explore further in Chapter 3 of this thesis. However, we need not assume this has to be the case: fitness payoffs may instead be taken to refer to some other mechanism through which preference types reproduce.[11]

A difficulty for theories of preference evolution that appeal to observability of types is that they are a form of "green beard" explanation.[12] Imagine that some individuals inherit a visible characteristic, such as having a green beard, and are predisposed to cooperate with other green-bearded individuals. At first, they may enjoy a fitness advantage. Yet if nature can generate mutually-cooperating types with green beards, there is no obvious reason why it should not generate materialists with green beards, in effect making preferences unobservable. However, there may in fact be a barrier to such imitation by materialists. For example, suppose that observing a preference type does not involve observing a fixed visible trait, but rather an inference made on the basis of interacting with others, e.g. through conversation. In this case, observing preference types may be costly, but so too may be imitating a preference type that is not one's own. With this environment in mind, the green beard problem may be a reason for extending observable-type models by making type observation a costly activity, as in my model. The alternative approach would be to allow players to imitate types, as in Wiseman and Yilankaya (2001), in whose model the population shares of types are cyclical. Their work builds on that of Robson (1990).

In section 1.2, I present the model. In section 1.3, I define equilibrium play and impose assumptions to select a unique equilibrium, which I then analyse. Section 1.4 deals with the evolution of preferences given the unique equilibrium. Section 1.5 concludes the chapter.

[11] Dennett (1995) argues that cultural evolution and genetic evolution can be thought of as instances of the same phenomenon; the mechanism through which cultural evolution takes place is much less readily identified, however. A recent account of such a mechanism can be found in Boyd and Richerson (2009).

[12] The term "green beard" is due to Dawkins (1976).

1.2 The model

1.2.1 Outline

A continuum $[0, 1]$ of agents undergoes uniform random pairing to play a two-player symmetric game Γ with action set $\{c, d\}$. The game Γ thus has four outcomes: (c, c), (c, d), (d, c) and (d, d), where I adopt the labelling convention that the first entry is the action of player i and the second that of player $j \neq i$ where $i, j \in [0, 1]$. Denote an arbitrary outcome $\mathbf{z} = (z_i, z_j) \in \{c, d\}^2$. Each player has a preference type. The game Γ specifies a set of fitness payoffs over the four outcomes, and specifies a set of "subjective" payoffs – i.e. von Neumann-Morgernstern (vNM) payoffs – for each preference type over the four outcomes.

1.2.2 Fitness payoffs

The fitness payoffs of the game Γ form a PD, as in Figure 1, where if an agent plays c (i.e. she cooperates) then she incurs a cost (normalised to 1) in the form of a reduction in her own fitness, but in so doing gives a fitness benefit $b > 1$ to her opponent.

Figure 1: Structure of fitness payoffs forming PD

		Player 2	
		c	d
Player 1	c	*b-1 , b-1*	*-1 , b*
	d	*b , -1*	0 , 0

Playing d (i.e. defecting) is thus strictly dominant in fitness payoffs. Mutual cooperation (i.e. action profile (c, c)) is efficient in fitness payoffs, since total fitness for the two players is $2(b-1)$, which is greater that of $(b-1)$ from (c, d) or (d, c), which is in turn greater than the total fitness of zero from the mutual defection profile (d, d).

For example, setting $b = 3$ yields the table of fitness payoffs in Figure 2.

Figure 2: Example of fitness payoffs: $b = 3$

		Player 2	
		c	d
Player 1	c	2 , 2	-1 , 3
	d	3 , -1	0 , 0

1.2.3 Research choices

In the **model with technology** (the "full" model), before agents undergo uniform random pairing to play Γ, every player i makes a private "research choice" $r_i \in \{\underline{p}, \overline{p}\}$, where the *baseline probability* of type revelation $\underline{p} \in (0, \overline{p})$ and $\overline{p} \leq 1$. If $r_i = \overline{p}$ we say that i *does research*. The research choice r_i is the probability with which i learns her opponent j's preference type. Define $\triangle p := \overline{p} - \underline{p} \in (0, 1)$. In other words, $\triangle p$ is the additional probability an opponent's type is revealed due to doing research.

The probability r_i that i observes her opponent j's type is independent of whether j observes i's preference type. Research choice r_i imposes a fitness a cost $k(r_i)$ on i, where:

$$k(r_i) = \begin{cases} 0 & r_i = \underline{p} \\ k & r_i = \overline{p} \end{cases} \tag{1.1}$$

where, in turn, $k > 0$ is a constant. In the **no-technology model**, players do not make a research choice. Instead, they simply undergo uniform random pairing whereupon every player i learns her opponent j's preference type with probability $\underline{p}$, before playing Γ.

1.2.4 Preference types

All players in the full model have preferences over the research cost $k(\cdot)$ and the four possible outcomes of the game Γ.[13] These preferences can be represented by vNM utilities, called

[13]In the no-technology model, all players have preferences over the four possible outcomes of the game Γ. I omit definitions of utility functions and strategies for the no-technology model as these are analogous to those

"subjective payoffs". For all agents, subjective payoffs from outcomes (c,d), (d,c) and (d,d) are equal to the fitness payoffs, respectively -1, b and 0; the subjective payoff from the research cost is fixed to be the fitness payoff $-k(\cdot)$.[14] Player i has utility function $u_i : \{c,d\}^2 \times [0,1] \to \mathbb{R}$ from outcomes together with research costs to the reals.

Preference types are characterised by the remaining subjective payoff, that from outcome (c,c). One preference type, known as the *materialist* preference type, is denoted M. All other preference types, known as *reciprocator* preference types, are denoted θ, where $\theta b - 1$ is the agent's subjective payoff from (c,c). Let θ be a (reciprocator) preference type of player i. Then her subjective payoffs can be written

$$u_i(r,\mathbf{z}) = u_\theta(r,\mathbf{z}) = -k(r) + \begin{cases} \theta b - 1 & if\ \mathbf{z} = (c,c) \\ -1 & if\ \mathbf{z} = (c,d) \\ b & if\ \mathbf{z} = (d,c) \\ 0 & if\ \mathbf{z} = (d,d) \end{cases} \tag{1.2}$$

where $\theta \geq 1 + \frac{1}{b\underline{p}}$.[15] The subjective payoffs of a materialist (type M) player can be written

$$u_i(r,\mathbf{z}) = u_M(r,\mathbf{z}) = -k(r) + \begin{cases} b - 1 & if\ \mathbf{z} = (c,c) \\ -1 & if\ \mathbf{z} = (c,d) \\ b & if\ \mathbf{z} = (d,c) \\ 0 & if\ \mathbf{z} = (d,d) \end{cases} \tag{1.3}$$

in the full model, and the no-technology model is intended simply as an expositional aid in presenting the results of the full model.

[14] Agents are thus risk-neutral with respect to gambles over research cost.

[15] While the condition $\theta \geq 1 + \frac{1}{b}$ is sufficient to ensure that a reciprocator finds c to be a best response to an opponent's playing c, I adopt the condition $\theta \geq 1 + \frac{1}{b\underline{p}}$ to ensure there is an equilibrium in which cooperation takes place at arbitrarily small reciprocator population shares. This in turn motivates the simple metric of attainability that I use to characterise the evolutionary implications of the discovery technology. Specifically, it permits the interpretation of attainability as the maximum population share which an arbitrarily small invasion of reciprocators into a materialist population will attain under payoff-monotone fitness dynamics.

Reciprocators are thus defined to be all players whose subjective payoff from (c, c) is at least as great as that from (d, c). The set of all reciprocator types is denoted $\Theta(b, \underline{p}) \equiv [1 + \frac{1}{b\underline{p}}, \infty)$, with arbitrary element denoted $\theta \in \Theta(b, \underline{p})$.

Figure 3: Structure of subjective payoffs for two players of type $\theta \in \Theta(b, \underline{p})$

		Player 2	
		c	d
Player 1	c	$\theta b - 1$, $\theta b - 1$	*-1* , *b*
	d	*b* , *-1*	0 , 0

Figure 4: Example of subjective payoffs for two players of type $\theta = 2$, where $b = 3$

		Player 2	
		c	d
Player 1	c	5 , 5	-1 , 3
	d	3 , -1	0 , 0

If two reciprocators (i.e. two players with types $\theta \geq 1 + \frac{1}{b\underline{p}}$) meet, the game they play is therefore a coordination game in subjective payoffs.

I focus solely on distributions that contain at most two types, which are in general denoted (θ, M, ε), where $\theta \geq 1 + \frac{1}{b\underline{p}}$ is the first type (a reciprocator type), M is the second type (the materialist type) and $\varepsilon \in [0, 1]$ is the population share of reciprocators. Recalling that the set of players is modelled as the unit interval $[0, 1]$, let $[0, \varepsilon]$ represent those players of type θ – the reciprocators – and let $(\varepsilon, 1]$ represent those players of type M, the materialists.

1.2.5 Strategies

Fix a reciprocator type $\theta \geq 1 + \frac{1}{b\underline{p}}$. Denote player i 's strategy $(r_i(\varepsilon), \mathbf{a}_i(\varepsilon))$. Denote a strategy profile $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)) = (r_i(\varepsilon), \mathbf{r}_{-i}(\varepsilon); \mathbf{a}_i(\varepsilon), \mathbf{a}_{-i}(\varepsilon))$, where $-i$ denotes all players other than i, so that $(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}_{-i}(\varepsilon))$ is the profile of strategies of all players other than i. Player i's strategy has two stages.

Stage 1 strategy: research choice before playing Γ

The first stage $r_i(\varepsilon)$ is a mapping $[0,1] \longrightarrow \{\underline{p}, \overline{p}\}$ from type distributions to research choices. In other words, an agent makes her research choice $r_i(\varepsilon) \in \{\underline{p}, \overline{p}\}$ with knowledge of the type distribution, which is specified by $\varepsilon \in [0,1]$. A first stage strategy profile ("research profile") is denoted $\mathbf{r}(\varepsilon) = (r_i(\varepsilon), \mathbf{r}_{-i}(\varepsilon))$. Players make their research choices simultaneously and they do not observe other players' research choices.

Stage 2 strategy: conditional action in Γ

In stage 2, i has three information sets: either she learns her opponent is a reciprocator, she does not learn her opponent's type or she learns her opponent is a materialist.[16] Define player i's *conditional action vector* to be a triple $\mathbf{a}_i = (a_i^\theta, a_i^0, a_i^M) \in [0,1]^3$, where $a_i^\theta \in [0,1]$ is the probability with which i plays c (i.e. cooperates) after receiving information that the opponent is of (reciprocator) type θ, $a_i^0 \in [0,1]$ is the probability she plays c after receiving no information about the opponent's preference type and $a_i^M \in [0,1]$ is the probability she plays c after receiving information that the opponent is a materialist. The components of i's conditional action vector $\mathbf{a}_i$, i.e. a_i^θ, a_i^0 and a_i^M, are known as i's *conditional actions*. To save on notation, when denoting pure conditional action vectors I will simply write the triple $(1,0,1)$ as cdc, and so on.

The second stage of i's strategy, $\mathbf{a}_i(\varepsilon)$, is a mapping $[0,1] \longrightarrow [0,1]^3$ from type distributions to conditional action vectors.[17] The type distribution from which a player's opponent is

[16] In the model, players do not condition their second stage strategies on their first stage actions. Making this simplification does not change the set of distributions over outcomes in Γ that can be induced in equilibrium for a pair of players, compared with the alternative of allowing such conditioning. This can be proved simply, as follows. Suppose contrary to the model that players do in fact condition their second stage strategies on their first stage actions, so that stage 2 strategies are mappings $[0,1] \times [0,1] \longrightarrow \triangle\{[0,1]^3\}$ from first stage actions and type distributions to conditional actions. The equilibrium condition that applies to second stage strategies is condition 2. However, this condition is independent of i's first stage action (research choice) r_i. Consequently, if i knows she is off the equilibrium path in stage 2 (having made a research choice that is not played in equilibrium), her optimal continuation strategy is the same as it would be on the equilibrium path. Hence constraining stage 2 strategies to be mappings $[0,1] \longrightarrow \triangle\{[0,1]^3\}$ from type distributions to conditional actions does not restrict the set of equilibrium second stage conditional action profiles.

[17] In the no-technology model, this is i's entire strategy.

selected (under uniform random matching) is assumed to be the same for all agents independent of the agent's preference type and research choice. A profile of conditional action vectors – or *conditional action profile* – is denoted $\mathbf{a}(\varepsilon) \equiv (\mathbf{a}_i(\varepsilon), \mathbf{a}_{-i}(\varepsilon))$. Given such a profile, and fixing the preference type distribution (θ, M, ε), define

$$\mathbf{a}_\theta(\varepsilon) := \frac{1}{\varepsilon} \int_{i=0}^{\varepsilon} \mathbf{a}_i(\varepsilon)\, di \tag{1.4}$$

$\mathbf{a}_\theta(\varepsilon)$ the expectation of all the conditional action vectors played by reciprocators. This is because reciprocators are those players of type θ, whom we recall are indexed by $i \in [0, \varepsilon]$. The right hand side of (1.4) integrates over just those players in $[0, \varepsilon]$, with a normalisation factor of $1/\varepsilon$, and thus gives the expected conditional action among these players. If a player knows her opponent is a reciprocator, she believes her opponent's expected conditional action vector to be $\mathbf{a}_\theta(\varepsilon)$. As the vector $\mathbf{a}_\theta(\varepsilon)$ takes expectations over 3×1 vectors, it is itself a 3×1 vector. Its first element, $a^\theta_\theta(\varepsilon) \in [0, 1]$, is for example the expected probability with which c is played by a reciprocator conditional on meeting another reciprocator.

Similarly, define

$$\mathbf{a}_M(\varepsilon) := \frac{1}{1-\varepsilon} \int_{i=\varepsilon}^{1} \mathbf{a}_i(\varepsilon)\, di \tag{1.5}$$

$\mathbf{a}_M(\varepsilon)$ is the expectation of all the conditional action vectors played by materialists. Define

$$\mathbf{a}_0(\varepsilon) := \varepsilon \mathbf{a}_\theta(\varepsilon) + (1-\varepsilon)\mathbf{a}_M(\varepsilon) = \int_{i=0}^{1} \mathbf{a}_i(\varepsilon)\, di \tag{1.6}$$

$\mathbf{a}_0(\varepsilon)$ is thus the expected conditional action vector played by an opponent of unknown type given the conditional action profile $\mathbf{a}(\varepsilon)$. Given $\mathbf{a}_0(\varepsilon)$, the element $a^M_0(\varepsilon) \in [0, 1]$, for example, is the expected probability with which c is played by a randomly chosen player conditional on meeting a materialist.

Finally, define

$$r_\theta(\varepsilon) := \frac{1}{\varepsilon}\int_{i=0}^{\varepsilon} r_i(\varepsilon)di \tag{1.7}$$

and

$$r_M(\varepsilon) := \frac{1}{1-\varepsilon}\int_{i=\varepsilon}^{1} r_i(\varepsilon)di \tag{1.8}$$

$r_\theta(\varepsilon)$ and $r_M(\varepsilon)$ are the average research choices made by reciprocators and materialists respectively at $\varepsilon \in [0,1]$. By construction, for any $\varepsilon \in [0,1]$, $r_\theta(\varepsilon) \in [\underline{p}, \overline{p}]$ and $r_M(\varepsilon) \in [\underline{p}, \overline{p}]$.

When it comes to considering relative fitness in an evolutionary setting, I assume throughout (as is standard in the literature) that any learning happens in a short amount of time compared to the underlying evolutionary process, so that the fitness of any preference type in the population can be calculated on the basis that players' strategies are always in (Bayes-Nash) equilibrium.[18] Bayes-Nash equilibrium is the appropriate solution concept here, as opposed to Perfect Bayesian Equilibrium (PBE), due to two features of the game. The first feature is that players do not observe each other's actions in the first stage, and move simultaneously in the second stage, and so a player gains no additional information from the first stage with which she can update her beliefs about her opponent's type when she comes to play in the second stage. The second relevant feature of the game is that all players have a correct prior belief about the type distribution *ex ante*. These two features uniquely determine players' beliefs for each information set in the second stage. Consequently, the equilibrium condition of consistency for PBE would be redundant; PBE would not refine the set of Bayes-Nash equilibria in this game.

In order to define (Bayes-Nash) equilibrium, first, for any pair of conditional actions $(a_i, a_j) \in [0,1]^2$, define player i's *value* $v_i(a_i, a_j)$ as follows.

[18] Like the assumption that agents always know the distribution of types, the assumption that they play in equilibrium gives a tractable way of modelling optimal behaviour that has been learned on a short timescale in evolutionary terms.

$$v_i(a_i, a_j) := a_i a_j u_i(c,c) + a_i(1-a_j)u_i(c,d) + (1-a_i)a_j u_i(d,c) + (1-a_i)(1-a_j)u_i(d,d) \quad (1.9)$$

Player i's value is her interim expected utility when she and her opponent play conditional actions a_i and a_j respectively, i.e. cooperate with probabilities a_i and a_j respectively. The right hand side of (1.9) is the sum of i's expected utility for each outcome weighted by the probability of each outcome occurring in this situation. Note that $v_i(a_i, a_j)$ is strictly increasing in a_j for fixed a_i. This is a defining feature of a PD: whatever action i plays, she is better off in expectation the more likely j is to cooperate.

The definition of value via (1.9) can be used to characterise i's continuation value at each of her information sets, given a second stage strategy profile $\mathbf{a}(\varepsilon)$. Suppose i is a reciprocator, with type $\theta \in \Theta$.[19] Define i's continuation value on observing a reciprocator, $v_i^\theta(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$, as follows.

$$v_i^\theta(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)) \quad := \quad r_\theta(\varepsilon) v_i(a_i^\theta(\varepsilon), a_\theta^\theta(\varepsilon)) + (1 - r_\theta(\varepsilon)) v_i(a_i^\theta(\varepsilon), a_\theta^0(\varepsilon)) \qquad (1.10)$$

The first term on the right hand side of (1.10) is the product of r_θ – the expected probability with which j learns i's type – and i's value from playing $a_i^\theta(\varepsilon)$ when her reciprocator opponent j learns i is a reciprocator. The second term on the right hand side of (1.10) is the probability the reciprocator opponent does not learn i's type multiplied by i's value from playing $a_i^\theta(\varepsilon)$ in this situation. A notable property of $v_i^\theta(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ is that it is strictly increasing in $a_\theta^\theta(\varepsilon)$ and $a_\theta^0(\varepsilon)$ for fixed $r_\theta(\varepsilon)$ because $v_i(a_i, a_j)$ strictly increases in a_j. Player i's continuation value on observing a materialist, $v_i^M(\mathbf{a}(\varepsilon))$, is analogously defined.

$$v_i^M(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)) \quad := \quad r_M(\varepsilon)) v_i(a_i^M(\varepsilon), a_M^\theta(\varepsilon)) + (1 - r_M(\varepsilon)) v_i(a_i^M(\varepsilon), a_M^0(\varepsilon)) \qquad (1.11)$$

Finally, player i's continuation value when not learning her opponent's type, $v_i^0(\mathbf{a}(\varepsilon))$, is defined as follows.

[19] Definitions of continuation values in the case that i is a materialist type are closely analogous to those where she is a reciprocator, and are hence omitted for the sake of clarity.

$$\begin{aligned} v_i^0(\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)) \;:=\; & \varepsilon(r_\theta(\varepsilon) v_i(a_i^0(\varepsilon), a_\theta^\theta(\varepsilon)) + (1 - r_\theta(\varepsilon)) v_i(a_i^0(\varepsilon), a_\theta^0(\varepsilon)) \\ & + (1-\varepsilon)(r_M(\varepsilon) v_i(a_i^0(\varepsilon), a_M^\theta(\varepsilon)) + (1 - r_M(\varepsilon)) v_i(a_i^0(\varepsilon), a_M^0(\varepsilon)) \end{aligned} \tag{1.12}$$

The continuation value $v_i^0(\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ is the convex combination of i's value when unknowingly meeting a reciprocator, and that when unknowingly meeting a materialist, with the respective weights ε and $(1-\varepsilon)$ those of the population shares of reciprocators and materialists. Like the continuation value on observing a reciprocator $v_i^\theta(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$, the continuation value when not learning the opponent's type $v_i^0(\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ is strictly increasing in $a_\theta^\theta(\varepsilon)$ and $a_\theta^0(\varepsilon)$ for fixed $r_\theta(\varepsilon)$.

The expressions for i's value at each of her three information sets in stage 2 allow us to define player i's *ex ante expected utility* $Eu_i[\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)]$ induced at ε by $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$:

$$\begin{aligned} Eu_i[\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)] \;:=\; & r_i(\varepsilon)\varepsilon[v_i^\theta(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))] + r_i(\varepsilon)(1-\varepsilon)[v_i^M(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))] \\ & + (1 - r_i(\varepsilon)) v_i^0(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) - k(r_i(\varepsilon)) \end{aligned} \tag{1.13}$$

The first term on the right hand side of (1.13) is i's continuation value (given $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$) when she discovers that her opponent is a reciprocator, multiplied by $r_i(\varepsilon)\varepsilon$, the probability of reaching that information set (since $r_i(\varepsilon) \in \{\underline{p}, \overline{p}\}$ is the probability with which she discovers her opponent's type, and ε is the probability that her randomly matched opponent is a reciprocator). The second term is her continuation value of observing a materialist multiplied by the probability of reaching that information set, $r_i(\varepsilon)(1-\varepsilon)$. The third term on the right hand side of (1.13) is her continuation value when not learning her opponent's type, multiplied by $(1 - r_i(\varepsilon))$, the probability with which she does not learn her opponent's type. The final term arises from the cost of research.

Substituting the two possible values of $r_i(\varepsilon) \in \{\underline{p}, \overline{p}\}$ into (1.13) and taking the difference immediately yields an optimality condition with respect to i's research choice. Denote by $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$ i's *incentive to do research* at ε given strategy profile $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$, defined

as follows.

$$\begin{gathered}\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) := \\ \triangle p \left(\varepsilon v_i^{\theta}(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) + (1-\varepsilon) v_i^{M}(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) - v_i^{0}(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)\right) - k\end{gathered} \tag{1.14}$$

The first term on the right hand side of (1.14) represents the increase in continuation value given the conditional action profile in stage 2 as a result of changing the probability distribution over information sets due to doing research. Specifically, it is the product of (a) the increase in probability of observing the opponent's type from doing research, $\triangle p$, and (b) the difference between the expected continuation value when learning the opponent's type, $\varepsilon[v_i^{\theta}(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))] + (1-\varepsilon)[v_i^{M}(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))]$, and the continuation value when not learning the opponent's type, $v_i^{0}(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$. The second and final term is simply the cost of doing research. Player i finds it optimal to do research iff $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) \geq 0$. Her incentive to do research is increasing in her continuation values at the two information sets at which she learns her opponent's type and decreasing in the continuation value at the other information set.

A strategy profile $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)) \equiv (r_i(\varepsilon), \mathbf{r}_{-i}(\varepsilon), \mathbf{a}_i(\varepsilon), \mathbf{a}_{-i}(\varepsilon))$ is an *equilibrium* if, for any $\varepsilon \in [0,1]$, any player $i \in [0,1]$, for any conditional action vector $\mathbf{a}' \in [0,1]^3$ and for any research choice $r' \in \{\underline{p}, \overline{p}\}$, the following two inequalities hold.[20]

1. $Eu_i[\varepsilon, r_i(\varepsilon), \mathbf{r}_{-i}(\varepsilon), \mathbf{a}_i(\varepsilon), \mathbf{a}_{-i}(\varepsilon)] \geq Eu_i[\varepsilon, r', \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon), \mathbf{a}_{-i}(\varepsilon)]$

2. $Eu_i[\varepsilon, r_i(\varepsilon), \mathbf{r}_{-i}(\varepsilon), \mathbf{a}_i(\varepsilon), \mathbf{a}_{-i}(\varepsilon)] \geq Eu_i[\varepsilon, r_i(\varepsilon), \mathbf{r}_{-i}(\varepsilon), \mathbf{a}', \mathbf{a}_{-i}(\varepsilon)]$

Condition 1 requires that when players simultaneously make their research choices in stage 1, whatever the reciprocator population share $\varepsilon \in [0,1]$, the expected utility any player i receives from playing $r_i(\varepsilon)$ must be weakly greater than that from research choice $r' \in \{0,1\}$,

[20]The definition of equilibrium, and that of a strategy profile, is over all $\varepsilon \in [0,1]$. As equilibrium conditions 1 and 2 apply at each $\varepsilon \in [0,1]$ independently, I will in general talk about equilibrium selection and characterise equilibria *at a particular* $\varepsilon \in [0,1]$ or for intervals of ε, on the understanding that properly speaking, this identifies the class of all strategy profiles that meet the two equilibrium conditions and are as characterised at the particular value or interval of ε.

taking into account both the induced distribution over outcomes of Γ and research costs. In other words, given i's second stage strategy $\mathbf{a}_i(\varepsilon)$ and the other players' strategy profile $(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}_{-i}(\varepsilon))$, there cannot be any profitable deviation for i in terms of first stage research costs, whatever the preference type distribution. Condition 2 requires that i's expected utility be maximised by her conditional action vector in stage 2.[21]

Remark 1.1 A property of condition 2 that simplifies equilibrium analysis is that it is independent of i's own research choice $r_i(\varepsilon)$. The condition is equivalent to requiring that each component of $\mathbf{a}_i(\varepsilon)$ maximise i's expected utility at her respective information sets, which does not depend on her research choice. Because of this property, we know that if a player were off the equilibrium path after stage 1, it would still be optimal for her to play her stage 2 equilibrium strategy. In other words, condition 2 implies that on learning her opponent is a reciprocator, in expectation $a_i^{\theta}(\varepsilon)$ must be better than any alternative conditional action $a' \in [0,1]$, and similarly for $a_i^0(\varepsilon)$ and $a_i^M(\varepsilon)$ at their respective information sets.[22]

In general, there may be zero, one or many equilibrium strategy profiles in the second stage, depending on parameter values. As a consequence, in studying optimal behaviour in the next section, I deal with equilibrium selection. The equilibrium played is uniquely selected by requiring that reciprocators do maximal cooperation, along with two technical tie-breaking assumptions. In this equilibrium, all players of a type play the same pure strategy. Because the indirectly evolutionary framework involves a continuum of players, and because my model endogenises the probability of type revelation, which requires a sequential game, it takes some work to derive the unique equilibrium formally. Much of the material in subsections 1.3.1, 1.3.2 and 1.3.3 is devoted to this task, which I undertake for the sake of generality. Specifically, those subsections establish that a pure strategy equilibrium symmetric with respect to preference

[21] In the no-technology version of the model, the only play is in stage 2 and so equilibrium condition 1 does not apply. Equilibrium condition 2 then applies with fixed research choices $r_j = \underline{p}$ for all players $j \in [0,1]$.

[22] By the definition of *ex ante* expected utility in (1.13), the requirement in respective of $a_i^{\theta}(\varepsilon)$ is formally that for any $\varepsilon \in [0,1]$ and any $a' \in [0,1]$, $r_\theta v_i(a_i^{\theta}(\varepsilon), a_\theta^{\theta}(\varepsilon)) + (1 - r_\theta(\varepsilon)) v_i(a_i^{\theta}(\varepsilon), a_\theta^0(\varepsilon)) \geq r_\theta v_i(a', a_\theta^{\theta}(\varepsilon)) + (1 - r_\theta(\varepsilon)) v_i(a', a_\theta^0(\varepsilon))$, which, by inspection, is independent of i's research choice $r_i(\varepsilon)$.

type is played, which is the starting point for subsection 1.3.4.[23]

1.3 Equilibrium analysis

This section characterises equilibrium play in a population comprising reciprocators and materialists. Throughout, let a pair of preference types $\theta \in \Theta(b, \underline{p})$ and M be fixed.

1.3.1 Materialists' strategies

As d is the strictly dominant second-stage action for the materialist type, then for the reciprocator type, if i's opponent is revealed to i to be a materialist, i's optimal action is also to play d. It follows that a materialist's expected subjective payoff is decreasing in her research choice r, since research is costly; materialists can do no better than to set $r_M = \underline{p}$. This result is a specific instance of the following necessary condition, which holds for any type distribution.

Lemma 1.1 *In any equilibrium, if for any player i, $a_i^\theta(\varepsilon) = a_i^0(\varepsilon) = a_i^M(\varepsilon)$, then $r_i(\varepsilon) = \underline{p}$.*

Proof: The result follows straightforwardly from observing (i) that the opponent j's strategy is independent of r_i; (ii) i's optimal second stage action is independent of whether i learns j's type; and (iii) $\arg\min_{r \in \{\underline{p}, \overline{p}\}} \{k(r)\} = \underline{p}$, i.e. $r_i = \underline{p}$ uniquely minimises costs for i . □

Lemma 1.1 says that if a player's optimal second stage strategy is to play some fixed action regardless of information about her opponent's type, then her optimal research choice is zero. Research can only be valuable to an agent by raising the likelihood that an agent will learn her opponent's type before playing her action. If information does not influence an agent's optimal action choice then costly research cannot be optimal.

Recall that player i's conditional action vector $\mathbf{a}_i$ is a triple of second-stage actions, with the first element the action in case the opponent is revealed to be the reciprocator type θ,

[23] An alternative approach would be simply to assume that a pure strategy equilibrium symmetric with respect to preference type is played. Subsections 1.3.1, 1.3.2 and 1.3.3 establish that such an assumption is implied by Assumption 1.1 (which assumes maximal coordination by reciprocators).

the second element the action if the opponent's type is not revealed and the third element the action if the opponent is revealed to be type M. The triple *ddd* thus refers to the pure conditional action to play d whatever information there is on the opponent's type, for example. In similar fashion, *ccd* is the conditional action to play c (with probability 1) conditional on no information, c if the opponent is seen to be a reciprocator and d if one's opponent is seen to be a materialist.

From Lemma 1.1, we have the following result.[24]

Lemma 1.2 *Fix a pair of preference types $\theta \in \Theta(b, \underline{p})$ and M. Then, in any equilibrium, $r_M(\varepsilon) = 0$ and $\mathbf{a}_M(\varepsilon) = ddd$ for all $\varepsilon \in [0, 1]$.*

Proof: Fix an arbitrary value of $\varepsilon \in [0, 1]$. As d is a strictly dominant action for M in Γ by construction, $\mathbf{a}_M(\varepsilon) = ddd$ in any equilibrium. Lemma 1.1 then implies that $r_i = \underline{p}$ for all i with type θ, and so by (1.8), $r_M(\varepsilon) = \underline{p}$ for any $\varepsilon \in [0, 1]$. □

Having established materialists' equilibrium strategy, let us move on to reciprocators' strategies and equilibrium selection.[25]

1.3.2 Selecting among equilibria

Recall that reciprocator types are $\theta \geq 1 + \frac{1}{b\underline{p}}$. Reciprocators face a coordination problem, explained as follows. Take a pairing of two reciprocators, and suppose that each player knows their opponent is a reciprocator. Furthermore, suppose each believes it is highly likely that their own type has been revealed to their opponent. Even in this situation, however, a player may fear that her opponent does not trust her to play c, in which case she believes her opponent's rational response is to play d, and so her own rational action is also to play d.[26]

[24] An immediate consequence of Lemma 1.2 is that in any equilibrium $\mathbf{a}(\varepsilon)$, the continuation value on meeting a materialist $v_i^M(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)) = 0$, for any player $i \in [0, 1]$.

[25] Rather than prove the existence of equilibrium in general before discussing equilibrium selection, I wait until subsection 1.3.4 to demonstrate the existence of different symmetric pure-strategy equilibria.

[26] Failure of coordination due to mutual distrust, which becomes self-reinforcing, is sometimes known as "Schelling's dilemma", after Schelling (1960).

In order to rule out situations of this kind, first let us define the average *ex ante* expected utility among reciprocators $E[u_\theta(\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))]$ induced at population share ε by strategy profile $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$, as follows.

$$Eu_\theta[\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)] := \frac{1}{\varepsilon} \int_{i=0}^{\varepsilon} Eu_i[\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)]\, di \tag{1.15}$$

It will also be useful to define *i's incentive to cooperate given* a_j, $\triangle_i^c(a_j)$, as follows.

$$\triangle_i^c(a_j) := v_i(1, a_j) - v_i(0, a_j) \tag{1.16}$$

When facing an opponent j who plays a_j in expectation, i finds it optimal to cooperate iff $\triangle_i^c(a_j) \geq 0$. Substituting from (1.9) yields

$$\triangle_i^c(a_j) = a_j(u_i(c,c) - u_i(d,c)) + (1 - a_j)(u_i(c,d) - u_i(d,d)) \tag{1.17}$$

Recall that for all players, $u_i(d,d) > u_i(c,d)$. In addition, for reciprocators, by definition $u_i(c,c) \geq u_i(d,c)$. Equation (1.17) therefore establishes that the incentive to cooperate is strictly increasing in a_j if i is a reciprocator. This property is unsurprising; reciprocators rank mutual cooperation as the best outcome of Γ while they rank cooperating with a defector as the worst, so the return on cooperation is higher the more likely the opponent is to cooperate.

The following assumption ensures maximal cooperation among reciprocators.[27]

Assumption 1.1 *If* $Eu_\theta[\varepsilon, \mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon)] > Eu_\theta[(\varepsilon, \overline{\mathbf{r}}(\varepsilon), \overline{\mathbf{a}}(\varepsilon)]$ *where* $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ *and* $(\overline{\mathbf{r}}(\varepsilon), \overline{\mathbf{a}}(\varepsilon))$ *are equilibria, then* $(\overline{\mathbf{r}}(\varepsilon), \overline{\mathbf{a}}(\varepsilon))$ *is not played. Subject to this, if a player is indifferent between*

[27] I also make the (minor) assumption, to ensure a unique equilibrium is selected, that at $\varepsilon = 0$, if $\mathbf{a}_\theta(0) = cdd$ can be played in equilibrium, it is. The tie-breaking rule in Assumption 1.1 that indifferent players cooperate is needed to ensure that a unique (stage 2) conditional action profile is always selected, as the following example makes clear. Let $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ be an equilibrium in which all reciprocators cooperate at a given information set (e.g. conditional on no information about the opponent's type) at a certain value of ε, and suppose they are indifferent in doing so. Now consider a second-stage profile $\mathbf{a}'(\varepsilon)$ under which, at ε, a finite number of players defect for sure, but which otherwise is the same as $\mathbf{a}(\varepsilon)$. Then the new stage 2 strategy profile $\mathbf{a}'(\varepsilon)$ meets equilibrium condition 2, and as $\mathbf{a}_\theta(\varepsilon)$ (which takes *expectations* over the second stage strategy profile) is unchanged, equilibrium condition 1 still holds true for $\mathbf{r}(\varepsilon)$. In this case the equilibrium-constrained maximisation of a_θ^θ and a_θ^0 would not select between $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ and the latter equilibrium. The second tie-breaking rule, requiring that any player in indifferent between doing research or not in stage 1 do research, addresses an analogous issue, and is used in the proof of Lemma 1.4.

cooperating and not at any information set, then she cooperates. If a player is indifferent between doing research and not, she does research.

Assumption 1.1 says that reciprocators always coordinate with each other in stage 2 as much as it is possible for them to do so with no profitable unilateral deviations. Equivalently, this means that cooperation is maximised. To show this equivalence, let us take reciprocator i's strategy as fixed, and fix arbitrary $\varepsilon \in [0,1]$. Her *ex ante* expected utility is strictly increasing in the *ex ante* probability $r_\theta(\varepsilon)a^\theta_\theta(\varepsilon)+(1-r_\theta(\varepsilon))a^0_\theta(\varepsilon)$ with which she can expect a reciprocator opponent to cooperate (recalling that by Lemma 1.2, materialists never cooperate). It follows that if there are two equilibria $(\mathbf{r}(\varepsilon),\mathbf{a}(\varepsilon))$ and $(\overline{\mathbf{r}}(\varepsilon),\overline{\mathbf{a}}(\varepsilon))$ such that $Eu_\theta[\varepsilon,\mathbf{r}(\varepsilon),\mathbf{a}(\varepsilon)] > Eu_\theta[(\varepsilon,\overline{\mathbf{r}}(\varepsilon),\overline{\mathbf{a}}(\varepsilon)]$, then $r_\theta(\varepsilon)a^\theta_\theta(\varepsilon) + (1-r_\theta(\varepsilon))a^0_\theta(\varepsilon) > \overline{r}_\theta(\varepsilon)\overline{a}^\theta_\theta(\varepsilon) + (1-\overline{r}_\theta(\varepsilon))\overline{a}^0_\theta(\varepsilon)$. In other words, more cooperation happens in the former equilibrium. Assumption 1.1 yields the following Lemma.[28]

Lemma 1.3 *At any given $\varepsilon \in [0,1]$, either $\mathbf{a}_\theta(\varepsilon) = ddd$, $\mathbf{a}_\theta(\varepsilon) = cdd$ or $\mathbf{a}_\theta(\varepsilon) = ccd$.*

Proof: See Appendix A.1. The fact Assumption 1.1 implies that symmetric pure strategy profiles are played is not surprising given the *ex ante* symmetry of the two-stage game with respect to players of a given type. The fact that $a^\theta_i(\varepsilon) \geq a^0_i(\varepsilon)$ is also to be expected, as a player's continuation value at the information set where she does not learn her opponent's type is a convex combination of her continuation values at her other two information sets, and $v^M_i(\mathbf{a}(\varepsilon)) = 0$ in any equilibrium.

1.3.3 Constrained-optimal research choices for reciprocators

Having considered the incentive to cooperate, I now consider how an individual reciprocator's incentive to do research depends on the research choices that other reciprocators make. To this end, take the partial derivative of $\Delta^r_i(\varepsilon,\mathbf{r}_{-i}(\varepsilon),\mathbf{a}(\varepsilon))$ with respect to $r_\theta(\varepsilon)$ for fixed ε, $r_M(\varepsilon)$ and $\mathbf{a}(\varepsilon)$, as follows.

[28] Lemma 1.3 relies on the fact that all reciprocators are of a single preference type θ; it would not hold in the case of a distribution with support from two or more types in $\Theta(b,\underline{p})$.

$$\frac{\partial \Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))}{\partial r_\theta(\varepsilon)} = \varepsilon \triangle p \left([v(a_i^\theta, a_\theta^\theta) - v(a_i^\theta, a_\theta^0)] - [v(a_i^0, a_\theta^\theta)) - v(a_i^0, a_\theta^0))] \right) \tag{1.18}$$

To interpret (1.18), first recall from (1.14) that i's incentive to do research, net of the cost of research, can be expressed as the product of $\triangle p$ and the difference between the expected value if the opponent's type is known, $\varepsilon v_i^\theta(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) + (1-\varepsilon) v_i^M(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$, and that when it is unknown, $v_i^0(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$. Now consider a small increase in $r_\theta(\varepsilon)$, the average research done by reciprocators. This increases the chance that a reciprocator opponent observes i's type θ and hence that the opponent plays a_θ^θ rather than a_θ^0. Player i's value at the information set where she learns her opponent is a reciprocator, $v_i^\theta(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$, therefore increases in proportion to the difference $v(a_i^\theta, a_\theta^\theta) - v(a_i^\theta, a_\theta^0)$. In similar fashion, a small increase in $r_\theta(\varepsilon)$ increases i's value at the information set where she does not learn her opponent's type. Specifically, $v_i^0(\mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$ increases in proportion to the difference $v(a_i^0, a_\theta^\theta) - v(a_i^0, a_\theta^0)$ multiplied by the probability $\varepsilon \in [0,1]$ that i is paired with a reciprocator. The partial derivative $\frac{\partial \Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))}{\partial r_\theta(\varepsilon)}$ thus represents the increase in expected value when the opponent's type is known minus the increase in expected value when it is not known.

Assumption 1.1 yields the following result, which will enable a tractable characterisation of equilibria by different regions of parameter values.

Lemma 1.4 *For any share of reciprocators $\varepsilon \in [0,1]$, either $r_\theta(\varepsilon) = \underline{p}$ or $r_\theta(\varepsilon) = \overline{p}$.*

Proof: See Appendix A.1. Lemma 1.4 says that whichever equilibrium second strategy profile is played, either all reciprocators do research, or none do. The result follows from Assumption 1.1 and the fact that reciprocators' research choices are weak complements: if it is optimal for one reciprocator to do research given second stage strategies, then it is optimal for all of them to do research.

Remark 1.2 Assumption 1.1 ensures that the research profile played maximises the ex ante probability of mutual cooperation for a given second stage strategy profile, just as it ensures

that the second stage profile played maximises the probability of mutual cooperation. If $\mathbf{a}_\theta(\varepsilon) = ddd$, then cooperation never takes place, so any research profile trivially maximises cooperation. If $\mathbf{a}_\theta(\varepsilon) = ccd$, then the level of mutual cooperation is independent of research, since mutual cooperation can only happen when two reciprocators meet, and they will always cooperate when they meet regardless of whether they observe each other's type. Finally, if $\mathbf{a}_\theta(\varepsilon) = cdd$ then mutual cooperation only takes place if two reciprocators meet and observe each others' types. This happens with probability $r_\theta(\varepsilon)^2$, so selecting the research profile that maximises $r_\theta(\varepsilon)$ maximises the ex ante probability of mutual cooperation.

1.3.4 Characterising equilibria

I now show that the assumption reciprocators can coordinate their strategies determines a unique equilibrium, which I characterise. Recall that as all materialists do no research and play ddd, an equilibrium can be identified by specifying reciprocators' strategies only. Lemmas 1.3 and 1.4 together imply that any given reciprocator population share $\varepsilon \in [0,1]$, $\mathbf{a}_\theta(\varepsilon) = ddd$, $\mathbf{a}_\theta(\varepsilon) = cdd$ or $\mathbf{a}_\theta(\varepsilon) = ccd$ and $r_\theta(\varepsilon) = \underline{p}$ or $r_\theta(\varepsilon) = \overline{p}$. Lemma 1.1 rules out the possibility that $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) = (\overline{p}, ddd)$, so there are five possible equilibria $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon))$ to consider: (i) $(\underline{p}, ddd)$; (ii) $(\underline{p}, cdd)$; (iii) $(\overline{p}, cdd)$; (iv) $(\underline{p}, ccd)$; and (v) $(\overline{p}, ccd)$.[29]

Applying the selection criteria of Assumption 1.1 to the candidate equilibria determines a unique equilibrium, as summarised in Proposition 1.1.

Proposition 1.1

1. *With or without a technology, reciprocators blindly cooperate (i.e. play $\mathbf{a}_\theta(\varepsilon) = ccd$) iff their population share $\varepsilon \geq \frac{1}{(\theta-1)b}$; otherwise, they play $\mathbf{a}_\theta(\varepsilon) = cdd$.*

2. *With a technology, reciprocators do research iff $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$ and $\varepsilon \in [\varepsilon', \varepsilon''] \subset [0,1]$, where $\varepsilon'' \geq \frac{1}{(\theta-1)b}$ and $\varepsilon' < \frac{1}{(\theta-1)b}$.*

[29] Furthermore, by the tie-breaking rules in Assumption 1.1, if $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) = (\underline{p}, cdd)$ then the equilibrium played is the one in which *all* reciprocators (not just any set of reciprocators of measure 1) play $r(\varepsilon) = \underline{p}$ and $\mathbf{a}(\varepsilon) = cdd$, and so on for the other cases.

Proof: See Appendix A.1. Proposition 1.1(1) says that reciprocators blindly cooperate when their population share is large enough, and that the population share at which this happens – the *free-riding threshold* – is inversely proportional to $(\theta - 1)b$. It follows that the higher the benefit from one's opponent cooperating, the lower the free-riding threshold. Equally, the higher the subjective benefit from mutual cooperation, the lower the free-riding threshold. The free-riding threshold is the point at which $\varepsilon(\theta b - b)$, the expected net benefit from mutual cooperation with other reciprocators that arises if all reciprocators blindly cooperate, equals one, the gross cost of cooperation.

Notably, the free-riding threshold is independent of the cost of research k, the baseline probability of type revelation $\underline{p}$ and the probability of type revelation induced by research $\overline{p}$. In short, reciprocators' decision to blindly cooperate is unaffected by the discovery technology. The reason for this is that the decision to blindly cooperate relates to the information set at which a player does not learn her opponent's type, whereas the discovery technology determines how likely this information set is to be reached.

Proposition 1.1(2) says first that reciprocators do research if the *effective cost of research* $\frac{k}{\triangle p}$ – i.e. the cost of research k scaled by the extent $\triangle p$ to which research increases type revelation probability – is low enough. For very high reciprocator types, i.e. high values of θ, the effective cost of research $\frac{k}{\triangle p}$ can be almost equal to one while making research optimal. This is because very high types blindly cooperate at low population shares, in which case doing research reduces the expected chance of being free-ridden by almost $\triangle p$. In the case of low reciprocator types, in contrast, the effective cost of research must be lower, with the "correction term" equal to the ratio between the cost of cooperation (equal to one) and the expect net benefit of cooperation of $(\theta - 1)b$. Proposition 1.1(2) then says that research is done over a contiguous interior interval of population shares. This is because at high population shares, where reciprocators blindly cooperate, their incentive to do research decreases in their population share, while at low population shares, where reciprocators blindly defect, their incentive to do research increases in population share.

Remark 1.3 One way of interpreting the use of the discovery technology is that individuals interrogate each other or ask around to find out about a potential partner's reputation, thereby hoping to discern their character (i.e. whether they have pro-social preferences). An alternative interpretation is that individuals' use of the discovery technology is done by 'screening' opponents by social distance. Social distance is a concept from sociology (see Simmel, 1950; for an early attempt to measure social distance, see Bogardus, 1924). It is a symmetric relation between individuals or groups, comprising a collection of features observable or discernible to both entities such as language(s) spoken, gender, social class, ethnicity, religion and nationality. Under this interpretation, players can adopt a policy of refusing to interact with others unless they are 'close enough' socially, as this improves the probability with which they discover their opponent's preferences. The cost involved can therefore be thought of as a form of search cost.

Figure 5: Equilibrium played by reciprocators for different values of k and ε, where $\theta = 2$, $b = 2$, $\underline{p} = \frac{1}{2}$ and $\overline{p} = 1$.

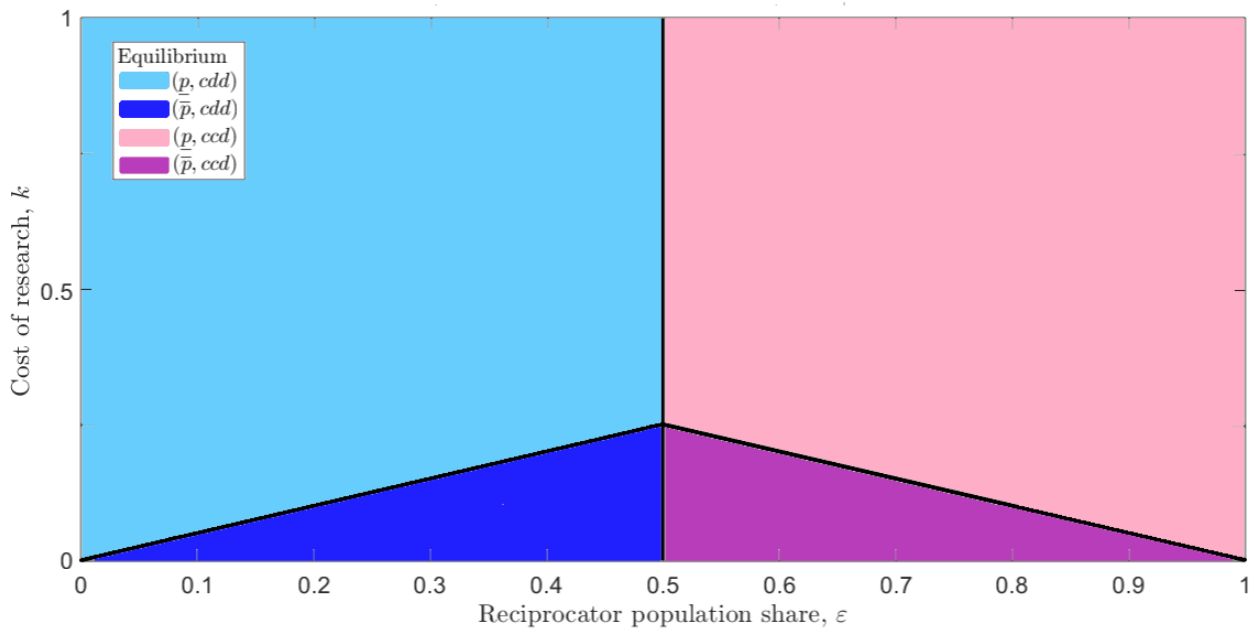


Figure 5 illustrates the equilibrium played for selected parameter values. Going from left

to right in the (ε, k) plane at a given height in the diagram traces increasing reciprocator population share ε, while descending represents decreasing the cost of research $\frac{k}{\triangle p}$. The vertical line at $\varepsilon = \frac{1}{2}$ is the free-riding threshold $\frac{1}{(\theta-1)b}$, to the right of which reciprocators blindly cooperate. The region formed by the triangle with bottom edge given by the horizontal axis is where research takes place; the incentive to do research where reciprocators blindly defect is increasing in reciprocator population share since research is more likely to yield cooperation the higher the value of ε. Reciprocators will therefore find it worthwhile to do research at higher costs as their population share increases to the left of the vertical line at $\varepsilon = \frac{1}{2}$. To the right of this line, the maximum tolerable research cost is downward-sloping in reciprocator population share, as the value of research lies in preventing being free-ridden by materialists.

Figure 6: Equilibrium played by reciprocators for different values of k and ε, where $\theta = 3$, $b = 2$, $\underline{p} = \frac{1}{2}$ and $\overline{p} = 1$.

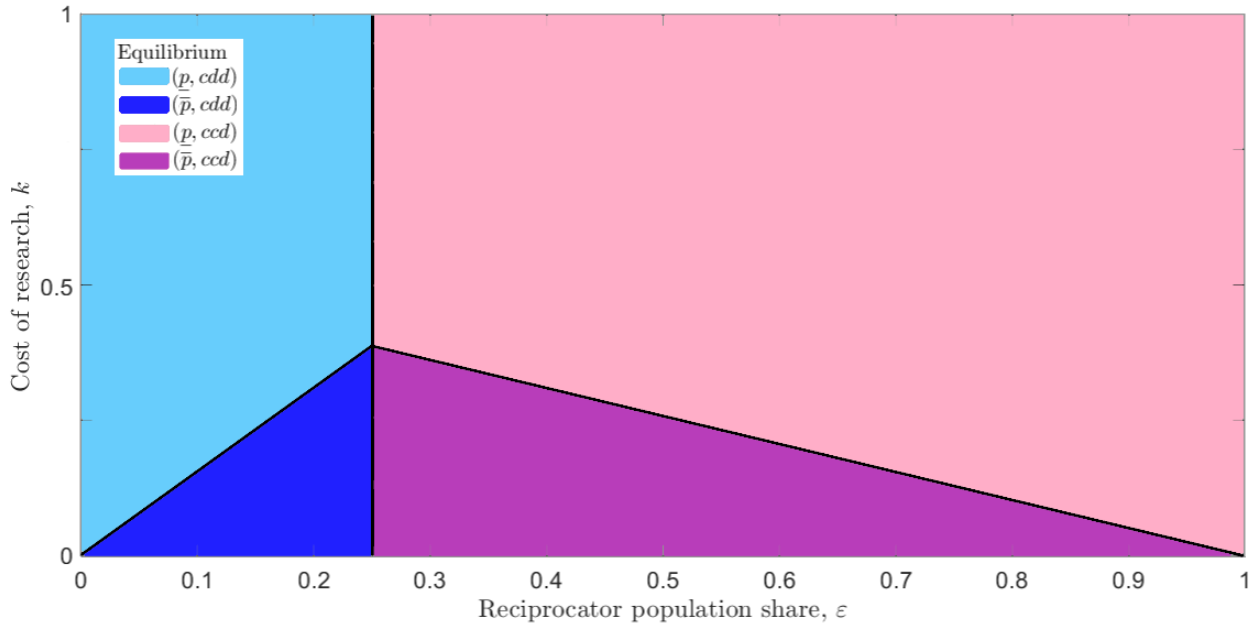


Figure 6 shows what happens for a higher reciprocator type, i.e. one which derives greater utility from mutual cooperation than the type in Figure 5. The free-riding threshold $\frac{1}{(\theta-1)b}$ is further to the left, as higher types find the additional mutual cooperation brought about by

blind cooperation more attractive for a given risk of being free-ridden. (Recall that materialists are only able to free-ride to the right of the vertical line at $\frac{1}{(\theta-1)b} = \frac{1}{4}$.) At the free-riding threshold, where the incentive to do research is strongest, reciprocators of this higher type ($\theta = 3$) tolerate higher research costs than do the lower type ($\theta = 2$) of Figure 5. Note that the upper boundary of the triangular region on the left hand side, in which reciprocators play $(\overline{p}, cdd)$, has a steeper upper boundary than before, because the higher type is willing to pay more to use the discovery technology. In contrast, the downward sloping line to the right of the vertical line, representing the maximum cost reciprocators are prepared to pay to use the technology when they blindly cooperate, has the same slope as in figure 5, since doing research in this region has the sole purpose of reducing the risk of being free-ridden, for which all types suffer the same disutility.

Figure 7: Equilibrium played by reciprocators for different values of k and ε, where $\theta = 2$, $b = 2$, $\underline{p} = \frac{1}{2}$ and $\overline{p} = \frac{3}{4}$.

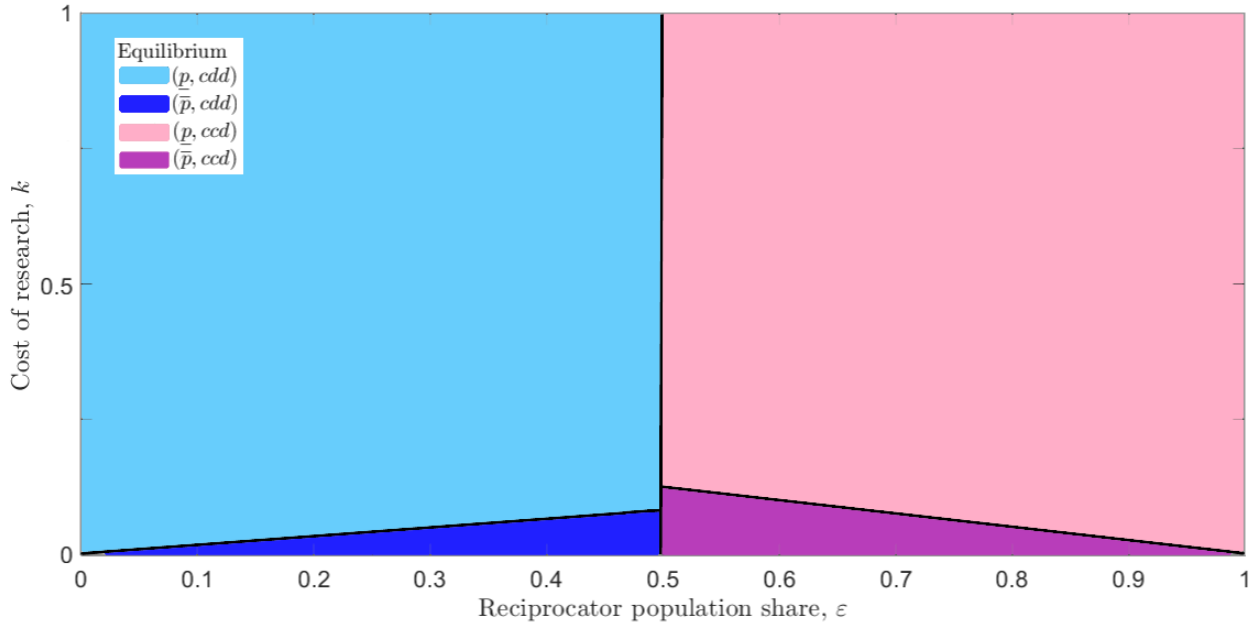


Figure 7 returns to the case that $\theta = 2$ but now considers what happens when the type discovery technology is less effective: specifically, $\underline{p} = \frac{1}{2}$ as before, but $\overline{p} = \frac{3}{4}$, less than before, so that $\triangle p = \frac{1}{4}$. In this case, one clear effect is that k, the cost of research, now represents

a lower effective cost of research $\frac{k}{\triangle p}$, and so reciprocators will not tolerate such high research costs as before. Another, more subtle, effect is that in the region where reciprocators blindly defect, their tolerance of research costs is reduced even further than in the region where they blindly cooperate. This is because their incentive to do research is driven by their subjective payoff from mutual cooperation, which is larger in magnitude than the disutility from being free-ridden that underpins the incentive to do research in the region on the right hand side. As such, their tolerance of costs is more sensitive to a reduction in $\overline{p}$ in the region on the left hand side.

Importantly, in the full model, the free-riding threshold $\frac{1}{(\theta-1)b}$ is independent of the choice of parameter values $k > 0$, $\underline{p} \in (0, \overline{p})$ and $\overline{p} \in (\underline{p}, 1)$. It is exactly the same as the free-riding threshold in the no-technology model. In the no-technology model, there is a simple reason why the free-riding threshold is independent of $\underline{p}$: if all reciprocators play *ccd* at some $\varepsilon \in [0, 1]$, then whether a reciprocator finds it optimal to deviate to playing $a_i^0(\varepsilon) = d$ (i.e. uninformed defection) depends only on her incentives when she does not learn her opponent's type. The probability of type revelation only determines how likely she is to reach this information set, not her incentives when she arrives at it.

However, in the full model, the fact $\frac{1}{(\theta-1)b}$ is independent of k, $\underline{p}$ and $\overline{p}$ is by no means obvious. Continuing to suppose all reciprocators play *ccd*, research is optimal iff $\varepsilon \leq 1 - \frac{k}{\triangle p}$. Equally, the values of k and $\triangle p$ determine whether doing research would be optimal were a reciprocator to deviate to play *cdd*. In other words, at the information set where a reciprocator does not learn her opponent's type, we need to take all the possible incentives to deviate into account, including those pure strategies that entail changing research choice. Indeed, at large reciprocator population shares research will not be valuable when playing *ccd* (since the risk of free-riding is low) but will be especially valuable in facilitating cooperation if deviating to *cdd*. Conversely, at low population shares, research will be highly valuable when playing *ccd* but not if deviating to *cdd*. While the incentives to do research are different at either side of the free-riding threshold, this difference does not affect the free-riding threshold itself.

1.4 Attainable type distributions

In this section, I introduce the concept of *attainability* to analyse how a discovery technology affects the evolution of a preference for reciprocity. I then characterise the attainability of different preference type distributions, given the unique equilibrium identified in Proposition 1.1. To keep things simple, I do not specify evolutionary dynamics, but rather I examine the maximum population share an arbitrarily small initial share of reciprocators will reach under payoff-monotone dynamics.

1.4.1 Definition of attainability

The indirect evolutionary approach typically involves identifying stable type distributions, i.e. those that are resilient against mutation, in some suitably-defined sense. Stability considerations also provide a basis for considering the dynamics of some specified type distribution of interest. Dekel, Ely and Yilankaya (2007) characterise a concept of *stable configuration*. Informally speaking, this is a distribution of preference types together with a Bayesian Nash equilibrium in which (i) all types present in the distribution receive the same fitness payoffs, and (ii) no small invasion of a new preference type can move the configuration "far away". Condition (ii) requires that the fitness of any new preference type not exceed that of the incumbents. It also requires that following an invasion, there is an equilibrium in which behaviour is not drastically different, in a sense made formal in the paper. As I have already imposed restrictions to select unique equilibria given any type distribution, and focus solely on two-type distributions, I am able to adopt a simpler notion of *attainability* that takes account of the particular form of equilibria characterised in the previous section.

To define this concept, first fix a distribution (M, θ, ε) and define by $F_\theta(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon))$ the expected fitness that a player of type θ receives in (M, θ, ε) given the uniquely selected equilibrium $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon))$ specified in Proposition 1.1 and given that materialists play $(\underline{p}, ddd)$.[30] Si-

[30] Algebraic expressions for $F_\theta(\varepsilon, \underline{p}, cdd)$, $F_\theta(\varepsilon, \overline{p}, ddd)$ and $F_\theta(\varepsilon, \underline{p}, ccd)$ and $F_\theta(\varepsilon, \overline{p}, ccd)$ are given in subsection 1.4.2.

milarly define by $F_M(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon))$ the expected fitness that a materialist receives in (M, θ, ε) given that *reciprocators* play $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon))$, and given that materialists play $(\underline{p}, ddd)$. Define the *relative fitness* $\triangle F(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon))$ as follows.

$$\triangle F(\theta, \varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) := F_\theta(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) - F_M(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) \tag{1.19}$$

Different reciprocator types can have different relative fitness because they may play different strategies, as reflected in $r_\theta(\varepsilon)$ and $\mathbf{a}_\theta(\varepsilon)$, which depend on θ. A positive relative fitness means that each reciprocator, on average, enjoys higher fitness than each materialist.

Let us temporarily fix $\theta \in \Theta(b, \underline{p})$. If we imagine reciprocators starting with an arbitrarily small population share, and the type distribution following dynamics that are monotone in fitness, such as the replicator equation, then in finite time reciprocators of type θ will reach their *attainable share* $\overline{\varepsilon}(\theta)$, defined as follows.[31]

$$\overline{\varepsilon}(\theta) := \sup_{\varepsilon \in [0,1]} \{\{0\} \cup \{\varepsilon \,:\, \forall \hat{\varepsilon} \in (0, \varepsilon)\,,\, \triangle F(\theta, \hat{\varepsilon}, r_\theta(\hat{\varepsilon}), \mathbf{a}_\theta(\hat{\varepsilon})) > 0\}\} \tag{1.20}$$

If we now allow θ to vary, from (1.20) we obtain a function $\overline{\varepsilon}(.) : \Theta(b, \underline{p}) \longrightarrow [0, 1]$ from reciprocator types to reciprocator population shares. For a given reciprocator type, $\overline{\varepsilon}(\theta)$ is the upper bound of the set containing zero and the contiguous region of population shares with infimum zero in which reciprocators enjoy strictly higher fitness than reciprocators.[32] Clearly, if population dynamics are such that $\triangle F(\theta, \varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) > 0$ implies $\frac{d\varepsilon}{dt} > 0$, and we suppose that initially $\varepsilon < \overline{\varepsilon}(\theta)$, the reciprocator population share will increase up to $\overline{\varepsilon}(\theta)$.

[31] The replicator equation applied to the model gives $\frac{d\varepsilon}{dt} = \varepsilon[F_\theta(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) - (\varepsilon F_\theta(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) + (1 - \varepsilon)F_\theta(\varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)))] = \varepsilon(1-\varepsilon)\triangle F(\theta, \varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon))$. Hence both at $\varepsilon = 0$ and at $\varepsilon = 1$, $\frac{d\varepsilon}{dt} = 0$, as one might expect. This is why the definition in (1.20) needs to use an interval which is open at the lower end and has an infimum of zero, rather than an interval containing zero. The reason it uses a supremum rather than a maximum is due to the fact the equilibrium $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ of Proposition 1.1 involves at least one discontinuous change in one element of $\mathbf{a}_\theta(\varepsilon)$, at the free riding threshold. It may also imply discontinuous changes in $r_\theta(\varepsilon)$ (if, for example, $\overline{p} < 1$, as in Figure 7). As a result, there is the possibility that $\triangle F(\theta, \varepsilon, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) > 0$ for all ε in an interval $[\varepsilon', \varepsilon'') \subset [0, 1]$, or an interval $(\varepsilon', \varepsilon'') \subset [0, 1]$, but that $\triangle F(\theta, \varepsilon, r_\theta(\varepsilon''), \mathbf{a}_\theta(\varepsilon'')) < 0$.

[32] Also, reciprocators never enjoy strictly higher fitness than reciprocators in the equilibrium played. Given this, the set over which the supremum is taken always includes zero, to ensure that $\overline{\varepsilon}(\theta)$ is always defined.

1.4.2 Expected fitness and attainable shares

Here I analyse the expected fitness for each type in the equilibrium identified in the previous section. First, I characterise relative fitness for each of the four pure strategy profiles that can be played by reciprocators in equilibrium, as follows.

- $\triangle F(\theta, \varepsilon, \underline{p}, cdd) = \varepsilon \underline{p}(b-1) > 0$
- $\triangle F(\theta, \varepsilon, \overline{p}, cdd) = \varepsilon \overline{p}(b-1) - k$
- $\triangle F(\theta, \varepsilon, \underline{p}, ccd) = \varepsilon \underline{p}(b-1) - (1-\underline{p})$
- $\triangle F(\theta, \varepsilon, \overline{p}, ccd) = \varepsilon \overline{p}(b-1) - (1-\overline{p}) - k$

The first two expressions are straightforward to understand; if reciprocators play $a^0_\theta = d$, then cooperation only arises in reciprocator-reciprocator pairs, and all materialists have zero fitness. Relative fitness is therefore given by the expected benefit $\varepsilon r_\theta(\varepsilon) b$ to a reciprocator from being paired with another reciprocator who observes her preference type (with probability $r_\theta(\varepsilon) \in \{\underline{p}, \overline{p}\}$), minus the expected cost of cooperation of $\varepsilon r_\theta(\varepsilon)$ (and, in the second expression, minus the cost of research k).

The last two expressions for relative fitness, for which $a^0_\theta = c$, require a little more interpretation because blind cooperation by reciprocators gives materialists non-zero expected fitness. Taking the expression for $\triangle F(\theta, \varepsilon, \underline{p}, ccd)$, for example, the first term on the right hand side arises from cooperation due to the conditional action $a^\theta_\theta = c$, as in the first two expressions. The second term on the right hand side arises from the fact that due to reciprocators' blind cooperation, which they do with probability $(1-\underline{p})$, materialists gain expected fitness of εb while reciprocators gain expected fitness of only $\varepsilon b - 1$.

One notable feature from comparing the relative fitness for the pure strategy profiles is that in each case, the relative fitness is strictly linearly increasing in the reciprocator population share ε, a property that can be explained as follows.[33] If $a^0_\theta = d$, relative fitness is simply

[33] This property proves useful in deriving results, because $\overline{\varepsilon}(\theta)$ can therefore be obtained by calculating the relative fitness at each lowermost population share for which a given symmetric pure strategy profile is played, and selecting the largest of these values such that it and all lower values induce positive relative fitness.

proportional to the probability ε that a reciprocator meets a fellow reciprocator, as materialists never get to free-ride. If $a^0_\theta = c$, then there will be a (negative) component of relative fitness proportional to $-(1-\varepsilon) - \varepsilon b = -1 - \varepsilon(b-1)$ resulting from free-riding, where the first term is the cost to reciprocators and the second is the benefit to materialists. Note, however, that the actual contribution to relative fitness from free-riding must be scaled down, because a materialist free-rides only if her type is not revealed. At the same time reciprocators' fitness from mutual cooperation is simply $\varepsilon(b-1)$. Hence the fitness difference is strictly linearly increasing in ε. The fact that type revelation reduces the scope for free-riding but not mutual cooperation also explains why reciprocators' fitness at $(\underline{p}, cdd)$ and $(\underline{p}, ccd)$ increases in $\underline{p}$ and that at $(\overline{p}, cdd)$ and $(\overline{p}, ccd)$ increases in $\overline{p}$. Furthermore, as the contribution of free-riding and mutual cooperation to relative fitness are both linear in $\varepsilon(b-1)$, the fact that free-riding happens less frequently because types are sometimes revealed explains why relative fitness is increasing in b.

The above characterisation of relative fitness at each of the pure symmetric strategy profiles played in equilibrium lays the groundwork for Lemma 1.5, which characterises the attainable share in the absence of a discovery technology. Its purpose is to provide a reference point for the main result of this chapter, in Theorem 1.1.

Lemma 1.5 *Fix a reciprocator type $\theta \in \Theta(b, \underline{p})$. Let $\overline{p} = 1$, so research reveals an opponent's type for sure, and fix $\underline{p} < 1$. Without a technology, if $\theta < 1 + \frac{\underline{p}(b-1)}{b(1-\underline{p})}$, then $\overline{\varepsilon}_{no\,tech}(\theta) = 1$ (reciprocators drive out materialists); otherwise $\overline{\varepsilon}_{no\,tech}(\theta) = \frac{1}{(\theta-1)b} \in (0,1)$.*

Proof: See Appendix A.1. Lemma 1.5 can be understood as follows. At low population shares, where reciprocators do blind defection, materialists cannot free-ride, and so reciprocators have a fitness advantage. At higher population shares, where reciprocators blindly cooperate, materialists are able to free ride, whenever they meet reciprocators who happen not to observe their preferences. Higher reciprocator types will switch to blind cooperation at a lower population threshold than will lower types, precisely because higher types derive greater utility from mutual cooperation. If the population threshold is too low – i.e. recipro-

cators of a given type switch too soon – then they will be prey to free-riding by materialists so often that they suffer lower fitness on average than materialists. The threshold type in this case is increasing in the probability of type revelation $\underline{p}$, since observability of types helps reciprocators; in the limit $\underline{p} \to 1$, any reciprocator type will drive out materialists from the population.

I now present the main result of this chapter, in Theorem 1.1. I set $\overline{p} = 1$, so that research reveals an opponent's type for sure: in other words, I assume a *perfect discovery technology*. This simplifies the statement of the theorem by equalising the maximum tolerable research cost in the region of population shares where reciprocators blindly defect to that in the region where they blindly cooperate.[34] Another motivation for setting $\overline{p} = 1$ is that it means if the technology is used, preferences are perfectly observable. Consequently, if the technology were free to use, it would enable any type of reciprocator to drive out materialists. This benchmark case is useful for comparison, allowing for ready interpretation of the results and clearly demonstrating the effect of introducing frictional costs into the model. For $\overline{p} < 1$ a weaker version of the main qualitative result holds, namely that depending on parameter values the discovery technology can raise or lower the attainable share, and that in the latter case, as $k \to 0$, $\overline{\varepsilon}(\theta) \to 0$. Figures 19 and 20 in Appendix A.2 provide examples for $\overline{p} = \frac{1}{2}$.

Theorem 1.1 *Assume a perfect discovery technology, i.e. fix* $\overline{p} = 1$ *and* $\underline{p} < 1$. *Fix a reciprocator type* θ *and let* $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$, *so reciprocators do research at some population share. Then* $\exists\, \theta' \in \Theta(b, \underline{p})$ *such that:*

1. *If* $\underline{p}(b+1) - \underline{p}^2 - 1 \leq 0$ *or* $\theta \geq \theta'$, *the technology reduces reciprocators' attainable share:* $\overline{\varepsilon}(\theta) < \overline{\varepsilon}_{no\,tech}(\theta)$. *Furthermore, as* $k \to 0$, $\overline{\varepsilon}(\theta) \to 0$.

2. *If* $\underline{p}(b+1) - \underline{p}^2 - 1 > 0$ *and* $\theta < \theta'$, *then* $\overline{\varepsilon}(\theta) \geq \overline{\varepsilon}_{no\,tech}(\theta)$.

[34] In other words, the technology is used in a region like the one one in Figures 5 or 6, rather than like the one in Figure 7 where $\overline{p} < 1$.

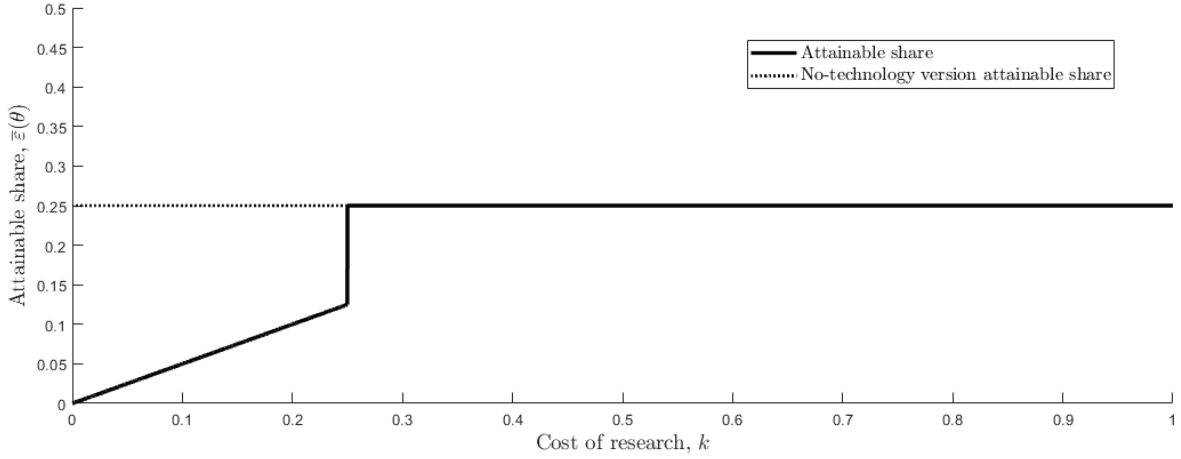
Proof: See Appendix A.1. Theorem 1.1 assumes a perfect discovery technology and assumes that $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$, which by Proposition 1.1 is necessary and sufficient for reciprocators to do research at some population share. Theorem 1.1(1) identifies the conditions under which reciprocators' attainable share is lower as a result of the technology. One sufficient condition is that the baseline probability of type revelation $\underline{p}$ is low enough: specifically $\underline{p}(b+1) - \underline{p}^2 - 1 \leq 0$. Note that the left hand side is an increasing function in $\underline{p}$ and in b over their joint domain; intuitively, the former improves observability of preferences, the channel through which a preference for reciprocity may evolve, while the latter makes mutual cooperation more evolutionarily advantageous. Note also that the first term on the left hand side involves the product of the two variables, which are clearly complements. Theorem 1.1(1) says that if these variables are not high enough, then reciprocators will do worse, in evolutionary terms, than materialists. Put another way, if the fitness benefit b of mutual cooperation is low relative to the baseline probability of type revelation, then a preference for reciprocity will not evolve.

Importantly, moreover, the final statement in Theorem 1.1(1) is that for a given value of b, an arbitrarily cheap research cost will yield an arbitrarily small attainable share for all types of reciprocator if $\underline{p}$ is low enough. This means that for small research costs to inhibit the evolution of a preference for reciprocity, the type discovery technology must be not only cheap but also *effective*, in that $\triangle p = 1 - \underline{p}$ must be large enough.[35] As discussed in section 1.3.4, at lower population shares where reciprocators blindly defect, they may lose out to materialists because their incentive to do research is misaligned with the fitness benefits it brings them. Specifically, reciprocators' incentive to use the costly technology is to facilitate mutual cooperation, which by definition they 'overvalue' in evolutionary terms.

[35]The fact that a low baseline probability of type revelation (specifically, $\underline{p}$ such that $\underline{p}(b+1) - \underline{p}^2 - 1 \leq 0$) guarantees all reciprocator types do worse when the technology is present is linked to another result: if $\overline{\varepsilon}_{no\,tech}(\theta) = 1$, then $\overline{\varepsilon}(\theta) = 1$. Intuitively, if in the absence of the technology reciprocators drive out materialists, then their fitness advantage is too strong to be undermined by overusing the costly technology when it is present. The intuition is that for technology to undermine the evolution of reciprocity, the type discovery technology must be not only cheap but also effective, which is qualitatively the same condition as for the result in Theorem 1.1(1) that $\overline{\varepsilon}(\theta) \to 0$ as $k \to 0$.

Figure 8 illustrates the possibility that for certain values of research cost k, the discovery technology reduces the attainable share of a reciprocator type, for example parameter values. As can be clearly seen by the upward-sloping solid line in the left of the diagram, reciprocators' attainable share is in fact increasing in research cost k and becomes very small close to zero. Intuitively, reciprocators find it optimal to do research at very low population shares if the research cost k is very low, for which they pay a fitness cost that exceeds the expected fitness benefit it brings. Notably, in this example this 'research trap' holds for all reciprocator types, even those with low intensity; it is straightforward to verify that the condition $\underline{p}(b+1)-\underline{p}^2-1 = -\frac{1}{9} \leq 0$ in Theorem 1.1(1) is met.

Figure 8: A research trap: attainable share $\overline{\varepsilon}(\theta)$ for variable research cost k, where $\underline{p} = \frac{1}{3}$, $\overline{p} = 1$, $b = 2$ and $\theta = 3$, together with attainable share in absence of technology, $\overline{\varepsilon}_{notech}(\theta)$.



To gain insight into how the equilibrium behaviour characterised in section 1.3.4 determines the attainable share, Figure 9 shows the relative fitness enjoyed by reciprocators at different population shares for each of the symmetric pure strategy profiles $(\underline{p}, cdd)$, $(\overline{p}, cdd)$, $(\underline{p}, ccd)$ and $(\overline{p}, ccd)$, represented by coloured dashed lines, where $\underline{p} = \frac{1}{3}$, $\overline{p} = 1$, $b = 2$ and $\theta = 3$ as in Figure 8. For the profiles in which research is done, k is set equal to 0.1. I choose this value simply because it is a research cost for which $\overline{\varepsilon}(\theta) < \overline{\varepsilon}_{notech}(\theta)$ given the other example parameter values above. The dotted black line gives the relative fitness in the model without a discovery technology, where every player is constrained to play $r = \underline{p}$ in the first stage.

Note the free-riding threshold at $\frac{1}{(\theta-1)b} = \frac{1}{4}$; this is where reciprocators switch from blind defection to blind cooperation, resulting in a drop in relative fitness as materialists' attempts to free-ride meet with some success. For all the symmetric pure strategy profiles, relative fitness is linearly increasing in reciprocator population share, as discussed above. The values of relative fitness for the two strategy profiles $(\overline{p}, cdd)$ and $(\overline{p}, cdd)$ in which $r = \overline{p}$ coincide; this is unique to the case $\overline{p} = 1$, as perfect type observability ensures that the information set where an opponent's type is not observed, which is the only point at which the two strategy profiles differ, is never reached. At small population shares, these profiles imply negative relative fitness because cooperation is very rare, yet reciprocators still bear the research cost of $k = 0.1$. Relative fitness for these two strategy profiles is steeper than for the two profiles in which $r_\theta = \underline{p}$, because r_θ gives the marginal fitness benefit to reciprocators arising from a small increase in their population share.

Relative fitness when $(\underline{p}, cdd)$ is played is proportional to reciprocators' population share; no free-riding takes place and no research costs are incurred, so relative fitness depends solely on the level of cooperation (which happens among reciprocators only). Finally, relative fitness when $(\underline{p}, ccd)$ is played, while increasing linearly in population share, is always negative for the parameters chosen. In particular, the baseline probability $\underline{p} = \frac{1}{3}$ results in too much free-riding by materialists for reciprocators to overcome.

Figure 9: Relative fitness by reciprocator population share in the absence of a technology where $\underline{p} = \frac{1}{3}$, $b = 2$ and $\theta = 3$, together with relative fitness for strategy profiles where $\overline{p} = 1$ and $k = 0.1$.

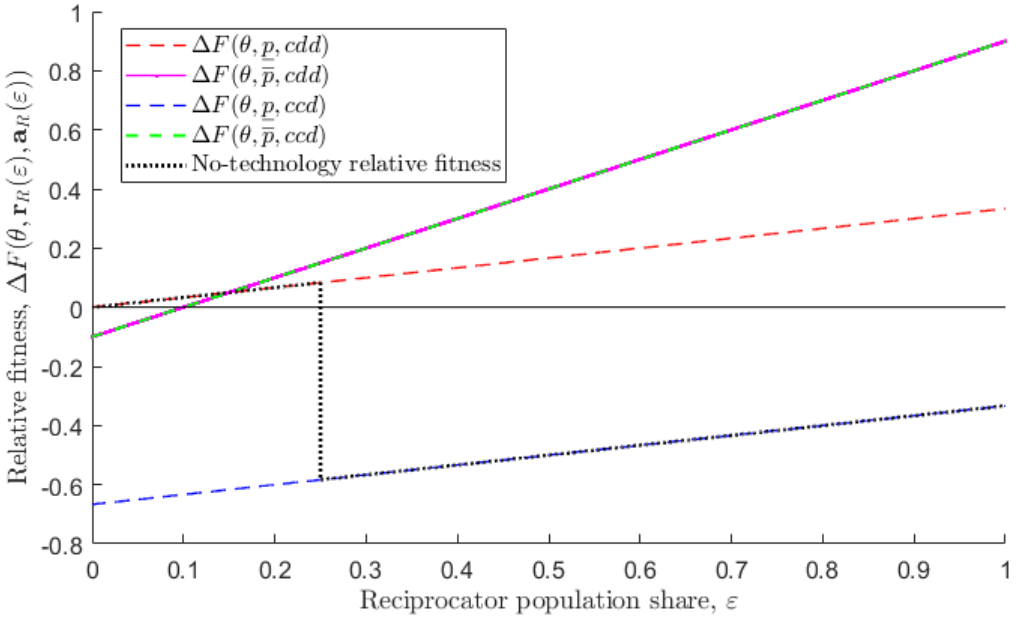
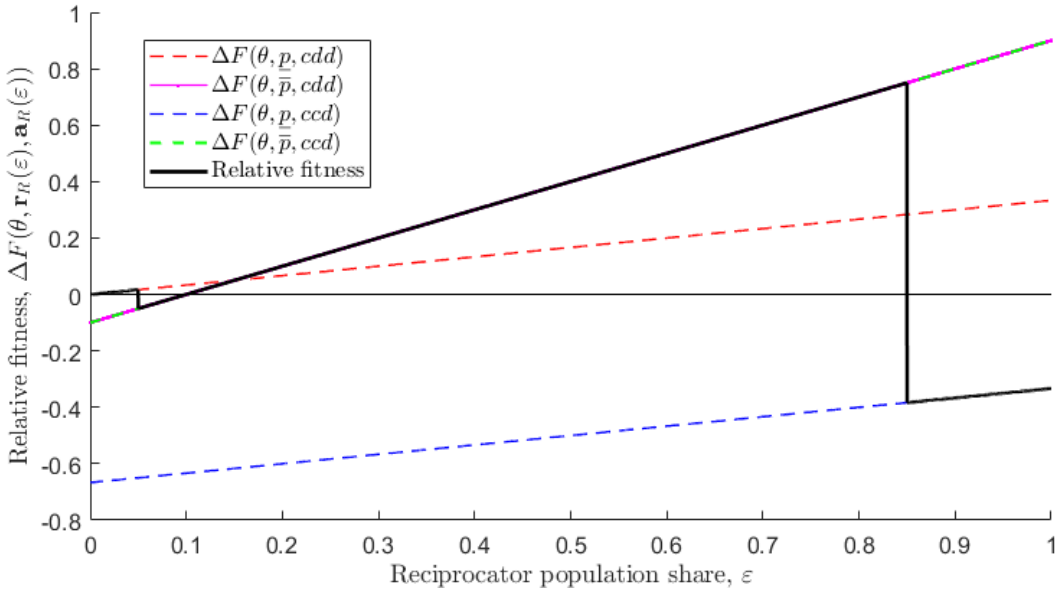


Figure 10 shows relative fitness when a discovery technology is introduced to the model with the parameters as in the previous two figures, including cost of research $k = 0.1$. The 'research trap' is clearly evident immediately to the left of the free-riding threshold. For lower values of k, this trap would move leftwards in the diagram, preventing an arbitrarily small share of reciprocators from growing to a large share of the population. Despite the fact that lower costs of information are in themselves better for reciprocators than higher such costs, reciprocators respond by incurring such costs at a lower population share, where their inherent fitness advantage over materialists is more slight.

Figure 10: Relative fitness by reciprocator population share with technology present, where $\underline{p} = \frac{1}{3}$, $b = 2$, $\theta = 3$, $\overline{p} = 1$ and $k = 0.1$



In summary, even small frictional costs can generate a research trap. Suppose a preference for reciprocity spontaneously arises (mutates) in a small number of individuals within a society where materialistic preferences are widespread. If preferences are freely observable, then on the rare occasion two reciprocators meet, they can collaborate on productive projects. However, if preferences are observable only at a small cost, reciprocators will bear this small cost many times, in many interactions with others. Most of the time, incurring the cost does not bear fruit, as reciprocators are rare in the society. The only exception is if the collaborative projects individuals can undertake together are highly productive, more than making up for the cumulative frictional costs.

1.5 Conclusion

In this chapter, I introduced a model that extends the literature on preference type evolution using the indirect evolutionary framework pioneered by Güth and Yaari (1992). I considered a population in which each player has one of two preference types: a reciprocator type –

which attaches greater utility to mutual cooperation than to free-riding – and a materialist type. The benchmark case for comparison is that under complete information, or if types are revealed with sufficiently high probability, then a preference for reciprocity will drive out materialistic preferences. In my model, such information is available to players via a "discovery technology", but only if they pay a cost, which reduces their evolutionary fitness.

I first studied the conditions under which players choose to bear a cost to improve their chances of learning their opponent's type. In general, reciprocators may willingly bear such a cost, but materialists never do. Reciprocators will do so only over a contiguous interior region of population shares. If the discovery technology is taken to represent 'screening' opponents by social distance, then this result implies that societies with intermediate levels of cooperation should be more fragmented into identity groups, whereas interactions between people in societies with either low or high levels of cooperation should be less conditioned on social distance. In contrast, another (more subtle) aspect of players' behaviour, cooperating when an opponent's preferences are not observed, is entirely unaffected by the discovery technology.

I then studied how costly information affects the evolution of reciprocity. Surprisingly, cheaper information can hinder the evolution of reciprocity more than expensive information, because it can tempt individuals that have a preference for reciprocity to 'overpay' for it. Indeed, if the fitness benefit from mutual cooperation is low, or the strength of preference for reciprocity is high, this misalignment of personal incentives and evolutionary interests means that even for arbitrarily cheap access to perfect observability of types, a preference for reciprocity cannot evolve (if it starts with a small share of the mix of preferences across the population). Intuitively, despite the fact that lower costs of information help reciprocators *ceteris paribus*, reciprocators respond by incurring such costs at a lower population share, where their inherent fitness advantage over materialists is more slight, resulting in a 'research trap'. This is a significant negative result for evolutionary explanations of reciprocity via preference type observability.

1.6 References

1. Alesina, A., Gennaioli, C. and S. Lovo, 2017. Public Goods and Ethnic Diversity: Evidence from Deforestation in Indonesia. National Bureau of Economic Research Working Paper no. 20504. Issued in September 2014, revised in January 2017.

2. Alexander, J., 2007. The Structural Evolution of Morality. Cambridge (UK): Cambridge University Press.

3. Alger, I. and J. Weibull, 2013. Homo Moralis: Preference Evolution under Incomplete Information and Assortative Matching. *Econometrica*, vol. 81, pp. 2269–2302.

4. Axelrod, R. and D. Hamilton, 1981. The Evolution of Cooperation. *Science*, vol. 211, no. 4489, pp. 1390-1396.

5. Becker, G., 1976. Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology. *Journal of Economic Literature*, vol. 14, pp. 817-826.

6. Bergstrom, T., 2003. The Algebra of Assortative Encounters and the Evolution of Cooperation. *International Game Theory Review*, vol. 5, pp. 211-228.

7. Binmore, K., 1998. Just Playing - Game Theory and the Social Contract, Part II. Cambridge (Massachusetts) and London: MIT Press.

8. Binmore, K. and A. Shaked, 2010. Experimental Economics: Where Next? Rejoinder. *Journal of Economic Behavior & Organization*, vol. 73, pp. 120-121.

9. Bogardus, E., 1925. Measuring Social Distances. *Journal of Applied Sociology*, vol. 9, pp. 299-308.

10. Bowles, S. and H. Gintis, 1998. The Evolution of Strong Reciprocity. Santa Fe Institute Working Paper, 98-08-073E 1998.

11. Boyd, R. and P. Richerson, 2009. Culture and the Evolution of Human Cooperation. *Philosophical Transactions of the Royal Society B*, vol. 364, no. 1533, pp. 3281-3288.

12. Charness, G. and M. Rabin, 2002. Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics*, vol. 117, pp. 817-869.

13. Dawkins, R., 1976. The Selfish Gene. Oxford: Oxford University Press.

14. Dekel, E., Ely, J. and O. Yilankaya, 2007. Evolution of Preferences. *Review of Economic Studies*, vol. 74, pp. 685-704.

15. Dennett, D., 1995. Darwin's Dangerous Idea. London: Penguin Books.

16. Esteban, J. and D. Ray, 2001. Collective Action and the Group Size Paradox. *American Political Science Review*, vol. 95, pp. 663-672.

17. Fehr, E. and A. Falk, 2002. Psychological Foundations of Incentives. European Economic Review, vol. 46, pp. 687-724.

18. Fehr, E., and K. Schmidt, 1999. A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics*, vol. 114, pp. 817-868.

19. Fudenberg, D. and E. Maskin, 1986. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica*, vol. 50, pp. 533–554.

20. Güth, W. and H. Kliemt, 1994. Competition or Cooperation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes. *Metroeconomica*, vol. 45 (2), pp. 155-187.

21. Güth, W. and M. Yaari, 1992. An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game. In *Explaining Process and Change Approaches to Evolutionary Economics*, ed. Witt, U., pp. 23-24. Ann Arbor: University of Michigan Press.

22. Guttman, J., 2000. On the evolutionary stability of preferences for reciprocity. *European Journal of Political Economy*, vol. 16, pp. 31–50.

23. Nowak, M. and R. May, 1992. Evolutionary Games and Spatial Chaos. *Nature*, vol. 359, pp. 826–829.

24. Ockenfels, P., 1993. Cooperation in Prisoners' Dilemma: An evolutionary approach. *European Journal of Political Economy*, vol. 9, no. 4, pp. 567-579.

25. Ok, E. and F. Vega-Redondo, 2001. On the Evolution of Individualistic Preferences: An Incomplete Information Scenario. *Journal of Economic Theory*, vol. 97, pp. 231-254.

26. Robson, A., 1990. Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake. *Journal of Theoretical Biology*, vol. 144, pp. 379-396.

27. Rohner, D., Thoenig, M. and and F. Zilbrotti, 2013. War Signals: A Theory of Trade, Trust, and Conflict. *Review of Economic Studies*, vol. 80, no. 3, pp. 1114-1147.

28. Schelling, T., 1960. Strategy of Conflict. Cambridge (Massachusetts): Harvard University Press.

29. Sethni, R. and E. Somnathan, 2001: Preference Evolution and Reciprocity. *Journal of Economic Theory*, vol. 97, pp. 273-297.

30. Simmel, G., 1950. The sociology of Georg Simmel. Ed. and trans., K. H. Wolff. Glencoe: Free Press.

31. Sobel, J., 2005. Interdependent Preferences and Reciprocity. *Journal of Economic Literature*, vol. XLIII (June), pp. 392–436.

32. Wiseman, O. and O. Yilankaya, 2001. Cooperation, Secret Handshakes, and Imitation in the Prisoners' Dilemma. *Games and Economic Behavior*, vol. 37(1), pp. 216-242.

Chapter 2 A theory of conditional cooperation on networks

Julien Gagnon and Alexander Harris[36]

2.1 Introduction

"*Reciprocity means that in response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self-interest model; conversely, in response to hostile actions they are frequently much more nasty...*" (Fehr and Gächter, 2000b: 159)

Reciprocity, or *conditional cooperation*, is ubiquitous in social interactions.[37] Mauss (1950) viewed reciprocity as "the human rock on which societies are built", while Gouldner famously presented the concept as one of the rare "universal principal components of moral codes" (Gouldner, 1960: 161). Likewise, Simmel (1950) deemed reciprocity necessary to cooperation and social cohesion in all societies, and regarded "all contacts among men [as resting] on the schema of giving and returning the equivalence" (Simmel, 1950; in Gouldner, 1960). In line with these views, current evidence unambiguously shows that most people display reciprocal inclinations, even with strangers, when it is costly to them or yields no future benefits.

[36]We first wish to thank Sanjeev Goyal and Robert Evans for continual support and comments. We also thank Francis Bloch, Antonio Cabrales, Vasco Carvalho, Matt Elliott, Aytek Erdil, Erik Eyster, Simon Gächter, Edoardo Gallo, Michael McBride, David Minarsch, Francesco Nava, Charles Roddie, Alex Wolitzky and seminar/conference participants in Cambridge, Chapman, Montreal, Nottingham, Paris, and Oxford for valuable comments. Julien Gagnon thanks the Gates Cambridge Trust and the Social Science and Humanities Research Council of Canada for financial support. Alex Harris is grateful for the financial support of the Economic and Social Research Council.

[37]We use the terms "reciprocity" and "conditional cooperation" interchangeably.

Social networks, through their ability to leverage social pressure and individuals' reciprocal inclinations, are seen as key to sustain cooperation in groups (see e.g. Granovetter, 2005; Burt, 1992; Coleman, 1990). However, social connections also make reciprocity fragile: while most individuals qualify as 'conditional cooperators', a few 'bad apples' (i.e. free-riders) are typically sufficient to influence others and derail cooperation in group interactions.[38] Given this tradeoff, under what circumstances will reciprocity-induced cooperation persist? What social architecture can best support it?

In this paper (i.e. Chapter 2 of this thesis), we argue that the structure of social interactions is key to the sustaining of reciprocity and cooperation in groups. We develop a model wherein connected agents can either contribute at a cost to a (local or global) public good or free-ride. Some players (*materialists*) care solely about material payoffs and always free-ride, while others (*conditional cooperators*) have *social payoffs*. Social payoffs capture the extent to which conditional cooperators feel influenced or pressured by the behaviour of players locally: their social payoffs from contributing increase with the number of their neighbours who contribute, and decrease with the number of their neighbours who free-ride. Social payoffs of this general form capture conditional cooperators' *reciprocal preferences*.[39]

We begin with a simple question: how can a network with both conditional cooperators and materialists support cooperation? A first observation is that since contribution is costly, any conditional cooperator i must have sufficiently high social payoffs to choose to cooperate. In particular, i must: one, be connected to enough other contributing players; and two, have a high enough proportion of her links with other contributing players. We find that at the unique maximal equilibrium (ME), a novel measure, the *q-linked set*, fully determines the set of players who cooperate (Theorem 2.1). The *q-linked set* consists in the largest set of conditional cooperators with both enough *cohesion* and enough *density*. The need for density

[38]Gächter (2006) shows that most experimental studies on reciprocity - despite differences in experimental design - yield a similar distribution of individual types, with about 55% of subjects being "conditional cooperators", 20% "free-riders", and the rest either "always contributors" or "triangle contributors".

[39]In Appendix B.2, we show that our model can be seen as a reduced form of an infinitely repeated local public good game where players are either forward-looking or myopic but have only material payoffs.

is driven by the (net) cost of contribution γ, which necessitates that enough *social pressure* be applied on conditional cooperators for them to cooperate. The need for cohesion is driven by conditional cooperators' reciprocal preferences, which requires enough positive *social influence* in their neighbourhood. We demonstrate that adding a link between a player in the γ-linked set and a player outside it can either increase or decrease cooperation (Proposition 2.1).

We then explore how a player's position in the network, for a given type profile, determines her *influence* on other players. We define a player i's influence as the proportion of players who are *susceptible* to i, i.e. players who would change their action at equilibrium if i's type changed. We show that a player i is susceptible to another player j if and only if she is connected to enough other players who are also susceptible to j (Proposition 2.2), and that susceptible players cannot be too interconnected (Corollary 2.1). Hence, a player i is influential *if she connected to a set of players interconnected enough* (allowing her influence to spread), *but not too much* (or her influence would be too diluted). Note that a player's influence and her centrality are not necessarily related.

Our analysis next leads us to examine how social interactions can be best structured to support reciprocity and cooperation. We consider a designer wishing to maximise cooperation and choosing the network at no cost. If the designer knew players' types, he could trivially group conditional cooperators and isolate materialists. Matters are more complex in the more realistic case where the designer only has the prior that each player's type is i.i.d. with *ex ante* probability p of being a materialist. We find that the *ex ante* optimal network is always formed by *isolated cliques* of degree $k^*(p)$, with $k^*(p)$ always above a threshold $\underline{k} \geq 1$ (Theorem 2.2). The intuition is that a network of cliques maximises the chances of clustering of conditional cooperators and minimises the extent and risk of contagion of free-riding.

We then study the effect of p on $k^*(p)$. To fix ideas, note that the designer faces a tradeoff when choosing k^*. On the one hand, increasing k^* entails the possible benefit of increasing the maximal number of materialists allowed in a clique for cooperation not to break down. On the other hand, increasing k^* entails the cost attached to an increasing number of 'draws'

and, thus, expected number of materialists in the clique. We find that $k^*(p)$ is decreasing in p (Proposition 2.4). This result is subtle: observe that increasing p yields two opposing effects on $k^*(p)$. First, a higher p raises the expected incidence of materialists. This effect pushes $k^*(p)$ *down*, as the designer wants to reduce the expected number of materialists in the clique. However, a higher p also reduces the expected incidence of conditional cooperators. This pushes $k^*(p)$ *up,* as the designer wants to increase the expected number of conditional cooperators in the clique. The former effect, it turns out, always dominates the latter.

Our model features a coordination game on networks and, as such, its results can be applied to a broad range of situations characterised by social interactions and behavioural contagion.[40] However, the model applies especially well to reciprocal social preferences, for three reasons. First, payoffs in the game are decomposed into "material" and "social" payoffs, with the latter capturing the defining features of reciprocal preferences. Second, contribution, unlike free-riding, is individually costly and entails positive externalities. Third, while conditional cooperators have coordination preferences, our model displays materialists who hamper coordination on the efficient outcome, in line with the empirical evidence on reciprocity.

We apply our results to one important application: work morale and peer influence at the workplace. Social preferences at the workplace are well-documented, and several models directly address social preferences and norms within firms (e.g. Kandel and Lazear, 1992; Huck et al., 2012; see Rotemberg, 2006, for a review). Our paper fills a gap in this literature by shedding light on a number of phenomena pertaining to networks and cooperation at the workplace. In particular, it explains how 'bad apples can spoil the barrel' by influencing workers around them, who in turn influence their own co-workers, and so on. It also explains why certain network structures, e.g. dense teams, are particularly vulnerable to such 'bad apples', while others are more robust by limiting their potential propagation. We discuss in

[40] In general terms, our model features a game of strategic complements on networks. Strategic complementarity has been the object of much study in the theory of supermodular games (see e.g. Topkis, 1979; Milgrom and Roberts, 1990; for a survey see Vives, 2005). Strategic complementary guarantees unique minimal and maximal equilibria in our game, as established by Topkis (1979). Our characterisation of the ME in terms of network structure and type profile is novel, and generalises results by Morris (2000) and Bollobas (1984).

detail in Section 2.6 the compelling evidence for our main results.

Related literature

Our paper is at the nexus of two major strands of research in economics. The first deals with social preferences, and more particularly with the role of *reciprocal preferences* in cooperation. There is a vast empirical literature establishing that most individuals display such preferences, i.e. they prefer to cooperate only when others do too (for surveys of the evidence, see e.g. Chaudhuri, 2011; Gächter, 2006; Fehr and Gächter, 2000b). Several well-established theories seek to capture reciprocal preferences by founding them either in distributional concerns (e.g. Charness and Rabin, 2002; Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) or in preferences over co-players' perceived intentions (e.g. Dufwenberg and Kirchsteiger, 2004; Rabin, 1993). Our model incorporates in a simple yet general way an essential feature of these models: players' inclination for cooperation decreases with the presence of free-riders, and increases with the presence of cooperators. To our knowledge, it is the first study of reciprocal preferences on networks.

The second main area of economics literature relevant here is games on networks (for recent surveys, see Bramoullé and Kranton, 2015; Jackson and Zénou, 2013). Recent papers address cooperation in repeated public good games on networks (Wolitzky, 2013; Ali and Miller, 2016, 2013; for a survey, see Nava, 2015). These papers extend earlier work on social sanctions, repeated interactions and cooperation (e.g. Ellison, 1994; Kandori, 1992). A common key finding of these papers is that denser networks, by allowing information (e.g. on defection) to better travel among players, foster cooperation. For this reason, central players tend to be the most cooperative. Our approach departs from these papers in several crucial ways. First, our model rests on reciprocal preferences, which enables us to account for some key stylised facts in line with the aforementioned empirical evidence.[41] Second, our model shows that network density may have adverse effects on cooperation if some players have reciprocal preferences.

[41]Reciprocity, as a behavioural trait, is prevalent even in absence of repeated interactions. Even in such settings, evidence suggests that the main driving force behind cooperation *and* social sanctioning is reciprocity (see e.g. Hopfensitz and Reuben, 2009; Fehr and Gächter, 2002, 2000a).

The reason is that denser networks allow bad behaviour, not just good behaviour, to propagate more easily. As a result, we show for instance that a player's influence on cooperation is not necessarily related to her centrality, and that the network that maximises ex ante contributions is not the complete network in general, as it is typically not robust to free-riders. Lastly, our model enables us to explore how to design networks to best support cooperation.[42]

Lastly, our paper closely relates to a large literature in the natural sciences on cooperation in structured societies. A major strand of this literature looks at the evolution of cooperation on fixed networks (for a survey, see e.g. Nowak, 2012). A key finding of these papers is that structured societies can permit some clustering of cooperative players, thereby sustaining some cooperation. In line with this work, recent experimental investigations have emphasised the role of conditional co-operators in sustaining cooperation on fixed networks, and have showed that "both cooperation and defection were contagious in fixed networks" (Jordan et al., 2013: 7) in non-repeated games among strangers (Rand et al., 2014; Jordan et al., 2013; Gracia-Lazaro et al., 2012; Suri and Watts, 2011; Fowler and Christakis, 2010; Grujic et al., 2010).[43,44] To our knowledge, our paper is the first to provide a formal framework of conditional cooperation on general fixed graphs. It allows us to formalise intuition regarding the mechanisms at play in these findings. It is also the first to formally investigate optimal network design in that context.

The rest of this chapter is divided as follows. Section 2.2 introduces the model. Section 2.4 explores the concepts of susceptibility and influence. Section 2.5 presents the study of optimal network design. We discuss our main application in Section 2.6. Section 2.7 concludes.

[42]Our results echo those of Haag and Lagunoff (2006) who study a repeated local prisoner's dilemma on networks where agents display heterogeneous discount factors. Our model and extension (Appendix B.2) allow for a wider class of preferences and payoffs. Our results also relate to those of recent papers on network formation and financial contagion (e.g. Cabrales et al., 2016; Erol and Vohra, 2014). Erol and Vohra (2014), in particular, build a model where forming links yield expected benefits to agents but also higher vulnerability. There is a similar tradeoff in our model. However, the authors here assume an exogenous condition for defaulting ("strong contagion"), i.e. a single agent defaulting entails that all nodes in her component default.

[43]Recent studies have also explored the role of endogenous network formation in sustaining cooperation (e.g. Gallo and Yan, 2015; Jordan et al., 2013; Rand et al., 2011; Fu et al., 2008).

[44]In contrast, the experimental literature in economics on the effect of network structure on cooperation has to date been modest (Falk et al, 2013 ; Carpenter et al., 2012 ; and Cassar, 2007, are a few notable exceptions).

2.2 The model

In this section, we develop a model to investigate how network structure influences reciprocity and cooperation in the context of public good provision. In particular, we explore the effect of having 'bad apples' in a group or team, i.e. players who never cooperate and influence their neighbours towards free-riding.

Network. Let $N = \{1, 2, ...n\}$ be the set of players, with $n \geq 3$. Denote by G an (undirected and unweighted) network, with a row vector $\mathbf{g}_i = \{g_{i1}, g_{i2}, ... g_{in}\}$ where $g_{ij} \in \{0, 1\}$ for all j, $i \in N$. Players i and j are *connected* iff $g_{ij} = 1$, and we assume $g_{ii} = 1$. Define by $N_i(G) = \{j \in N : j \neq i \wedge g_{ij} = 1\}$ player i's *neighbourhood*, and let $k_i = |N_i(G)|$ be her *degree*. There exists a *path* between players i and j either if $g_{ij} = 1$ or if there exists a set of players $\{i_1, i_2, ...i_j\}$ such that $g_{ii_1} = g_{i_1 i_2} = ... = g_{i_\ell j} = 1$.

Game and actions. Players interact in a one-shot public good game (PGG). Player i has action set $X_i = \{0, 1\}$. We denote i's action by $x_i \in X_i$ and the action profile of all players by $\mathbf{x} \in X = \{0, 1\}^n$. If $x_i = 1$ ($x_i = 0$), we say that i *contributes* (*free-rides*). The *contribution level* for a given action profile $\mathbf{x}$ is $\sum_{i \in N} x_i$. For a given $\mathbf{x}$, player i's *local action profile* is denoted by $\mathbf{x}_{j \in N_i(\mathbf{G})}$, and we denote by c_i and d_i, respectively, the number of i's neighbours who contribute and free-ride:

$$c_i = c_i\left(\mathbf{x}_{j \in N_i(G)}\right) := \sum_{j \in N_i(G)} x_j \tag{2.1}$$

$$d_i = d_i\left(\mathbf{x}_{j \in N_i(G)}\right) := \sum_{j \in N_i(G)} (1 - x_j) = k_i - c_i \tag{2.2}$$

Payoffs. Player i's payoffs are given by:

$$\pi_i(\mathbf{x}|\theta_i) = \sum_{j \in N \setminus \{i\}} x_j - \gamma x_i + \Psi(x_i, c_i, d_i|\theta_i) \tag{2.3}$$

where $\gamma > 0$ is the (net) *cost of contribution*.[45] Player i's type $\theta_i \in \Theta_i = \{\theta_M, \theta_R\}$ is ascribed

[45]Since the PG is assumed to be linear, all our results would hold even if it were a local PG. Further, while

by nature at the beginning of the game, with θ_M and θ_R referring respectively to *materialist* and *reciprocator* (or *conditional cooperator*) types. A type profile is denoted $\theta \in \Theta$. The function $\Psi(\cdot|\theta_i)$ captures i's *social payoffs.*

Assumption 2.1 *Players' payoffs are given by (2.3), where:* ***(1)*** $\Psi(\cdot|\theta_M) = 0$*;* ***(2)*** $\Psi(0, c_i, d_i|\theta_R) = 0$*;* ***(3)*** $\Psi(x_i, 0, 0|\theta_R) = 0$*; and* ***(4)*** $\Psi(1, c_i, d_i|\theta_i)$ *is (weakly) increasing in* c_i *and is (weakly) decreasing in* d_i*.*

We highlight some key elements of Assumption 2.1. *First*, note that since $\gamma > 0$, Assumption 2.1(1) guarantees that free-riding is always a strictly dominant strategy for materialist players. *Second*, Assumption 2.1(2) states that players who free-ride always have zero social payoffs. This is a simplification for ease of exposition.[46] *Third*, we assume that isolated players have no social payoffs. In that respect, an isolated player is viewed as immune to any social influence/pressure. *Fourth*, Assumption 2.1(4) captures, in a general and parsimonious way, preferences for conditional cooperation and reciprocity: conditional cooperators' social payoffs from contributing increase (decrease) with the number of neighbouring contributors (free-riders). *Fifth*, note that we assume the social payoffs function to be the same for all conditional cooperators. All of our insights are robust to introducing heterogeneity in the social payoffs function, as long as it satisfies Assumption 2.1 for every conditional cooperator.

We now provide an example of a social payoffs function satisfying Assumption 2.1.

Example 2.1 Conditional cooperator i's social payoffs are given by:

$$\Psi(x_i, c_i, d_i|\theta_R) = x_i(\alpha c_i - \beta d_i) \tag{2.4}$$

with $\alpha > 0$ and $\beta > 0$.

we impose linearity, all of our results straightforwardly hold in the case of weakly convex global PG functions.

[46] Appendix B.2 relaxes Assumption 2.1(2). Furthermore, note that our insights, to hold, only require that cooperation exhibit strategic complementarity, which Assumption 2.1(2) and 2.1(3) together guarantee. For example, we could assume $\Psi(0, c_i, d_i|\theta_R)$ to be increasing (decreasing) in c_i (d_i), which would capture, *inter alia*, the increasing (decreasing) guilt of conditional cooperators when more neighbours contribute (free-ride).

Equation (2.4) offers a special case of a broad class of utility functions with interdependent preferences (see Sobel, 2005). Given our setup it is close to the reciprocity models offered in e.g. Charness and Rabin (2002) and Fehr and Schmidt (1999).

Equilibrium. We assume G and θ to be common knowledge among players.[47] Given G, θ and γ, an action profile $\mathbf{x} \in X$ is an *equilibrium* if for every $i \in N$ and every $x_i' \in X_i$, $\pi_i(x_i, \mathbf{x}_{-i}|\theta_i, G) \geq \pi_i(x_i', \mathbf{x}_{-i}|\theta_i, G)$. Observe that local complementarity in x entails potential coordination failures: in particular, the case where no player contributes is always an equilibrium. An equilibrium $\mathbf{x}^*$ is a *maximal equilibrium* (ME) if there does not exist another equilibrium $\mathbf{x}' \in \{0,1\}^n$ such that $\sum_{i \in N} x_i^* < \sum_{i \in N} x_i'$.

For the rest of this chapter, we restrict attention to ME. Three reasons motivate this focus. *First*, theoretical work on supermodular games shows that the ME is an upper bound on which play will converge for a very wide range of learning processes (Milgrom and Roberts, 1990). *Second*, the "all-free-ride" action profile is always an equilibrium for any network in our model; therefore, the ME implicitly characterises the range of equilibrium contribution levels (as it would if we allowed for mixed strategies in addition to pure ones). *Third*, the algorithm provided below, which ensures convergence to the ME, finds strong support in the experimental literature, which shows that conditional cooperators typically begin by cooperating before switching to free-riding if they observe too many free-riders.

2.3 Equilibrium characterisation

We now study the existence, uniqueness and properties of ME of this game. Consider a player i in a network G, and let q be some positive real number. Recall from Assumption 2.1 that $\Psi(1, c_i, d_i|\theta_R)$ is an increasing function of c_i and a decreasing function of d_i. Therefore, observe that either $\Psi(1, k_i, 0|\theta_R) < q$, or there exists a unique minimal $\overline{c}_i \in [1, k_i]$, such that:

[47] In Appendix B.3, we relax the assumption of knowledge of the type profile and assume that types are private and i.i.d. Our analysis yields a characterisation of a unique maximal Bayesian Nash equilibrium (BNE) analogous to that of the unique ME in our model. Our main results on comparative statics also hold in the case of this maximal BNE.

$$\Psi\left(1, \overline{c}_i, (k_i - \overline{c}_i)|\theta_R\right) = q \tag{2.5}$$

For any $q \in \mathbb{R}_+$, we define the *q-linked set* of G, denoted by $\mathcal{Q}_q(G) \subseteq N$, as the largest set of players such that for each $i \in \mathcal{Q}_q(G)$,

$$\Psi\left(1, s_i^q, (k_i - s_i^q)\right) \geq q \tag{2.6}$$

where $s_i^q :=\mid \{j \in N_i(G) \cap \mathcal{Q}_q(G)\} \mid$ is the number of i's neighbours in the set $\mathcal{Q}_q(G)$. Note that no materialist player can be in $\mathcal{Q}_q(G)$, as $\Psi(\cdot|\theta_M) = 0$. Note also that for arbitrary G and θ, the q-linked set is uniquely defined.

We provide an algorithm for the construction of the q-linked set of any network G, which we illustrate with Figure 11. Consider the payoff function (2.4), and let $\alpha = 1.25$ and $\beta = 1$. Suppose that $\theta_i = \theta_R$ for all $i \in N$. We find the 2.6-linked set.

Algorithm 1 *Fix initial profile* $\mathbf{x^0}$*, with* $x_i^0 = 1$ *for all* $i \in N$*. For a given* $q \in \mathbb{R}$*, assign* $x_i^1 = 0$ *for all* i *such that* $\Psi\left(1, c_i(\mathbf{x^0}), d_i(\mathbf{x^0})|\theta_i\right) < q$*, and denote the new profile by* $\mathbf{x^1}$*. Then, assign* $x_i^2 = 0$ *for all* i *such that* $\Psi_i\left(1, c_i(\mathbf{x^1}), d_i(\mathbf{x^1})|\theta_i\right) < q$*, and denote the new profile by* $\mathbf{x^2}$*. Iterate until step* k *where* $\mathbf{x^k} = \mathbf{x^{k+1}}$*. The nodes with* $x_i^k = 1$ *form the q-linked set.*

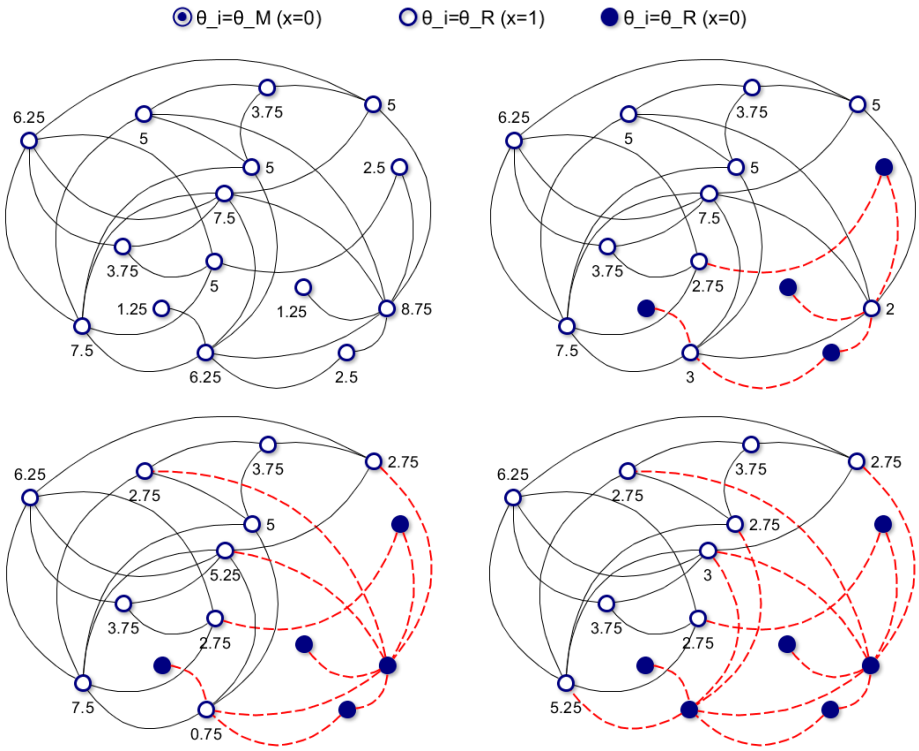
Theorem 2.1 *Suppose that Assumption 2.1 holds. For any* $\gamma \in \mathbb{R}_+$*,* G *and* θ*, a ME always exists and is unique. At the ME, a player contributes if and only if she is in the* γ*-linked set.*

Proof: All proofs for this chapter are in Appendix B.1.

Theorem 2.1 states that the decision to contribute for any conditional cooperator is uniquely determined by her belonging to the γ-linked set.[48] To contribute at the ME, a conditional cooperator must thus: one, be connected to *enough* other players in the γ-linked set; and two, have a *high enough proportion* of her links with other players in the γ-linked set. To see

[48]In any equilibrium, if a player cooperates, she must be in the γ-linked set. In particular, the set of cooperating players at any non-maximal equilibrium, say A, must be such that $A \subset \mathcal{Q}_\gamma$ and all players in A have enough links and in high enough proportion to other players in A.

Figure 11: The 2.6-linked set

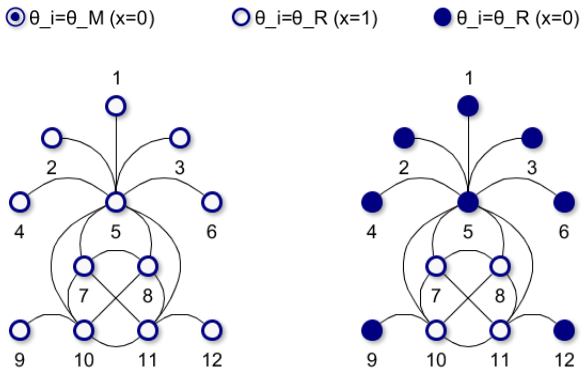


Top left: initial graph, with $\psi(\mathbf{x_0})$ for all nodes. **Top right**: Nodes with $\psi(\mathbf{x_0}) < 2.6$ are switched, and $\psi(\mathbf{x_1})$ are computed for remaining nodes. **Bottom left**: iteration. **Bottom right**: the 2.6-linked set obtains when no further iteration is possible.

why, consider again the payoffs function (2.4), and assume $\alpha = 1.5$ and $\beta = 1$. Suppose that $\theta_i = \theta_R$ for all $i \in N$, and fix $\gamma = 1.6$. Figure 12 illustrates the ME of a graph for these parameter values. At the ME $\mathbf{x}^*$, $x_i^* = 1$ if and only if $i \in \mathcal{Q}_{1.6} = \{7, 8, 10, 11\}$. Note that the proportion of neighbours of players 9 and 12 in $\mathcal{Q}_{1.6}$ is 100%: however, their total number of links to other players in $\mathcal{Q}_{1.6}$ is insufficient for them to be in $\mathcal{Q}_{1.6}$. Note also that player 5 has 4 links to players in $\mathcal{Q}_{1.6}$, higher than any other player. However, the proportion of her neighbours who are in $\mathcal{Q}_{1.6}$, namely $\frac{4}{9}$, is too low for her to be in $\mathcal{Q}_{1.6}$.

The concept of q-linked set is novel and combines the concepts of *q-core* (Bollobas, 1984) and *q-cohesive set* (Morris, 2000). The q-core is defined as the maximal set of players with at

Figure 12: Theorem 2.1 illustrated



Left: Initial graph. **Right**: the ME for $\alpha = 1.5$, $\beta = 1$, and $\gamma = 1.6$.

least q links to other players in the q-core. The q-cohesive set is the maximal set of players having a proportion of at least $q \in [0, 1]$ of their links with other players in the q-cohesive set. The q-linked set has a similar recursivity. However, a first critical difference is that the q-linked set necessitates both enough *cohesion* and *density*. On the one hand, the need for density is driven by the (net) cost of contribution γ, which requires that enough *social pressure* be applied on conditional cooperators to push them to cooperate. On the other hand, the need for cohesion is driven by conditional cooperators' reciprocal preferences, which necessitate enough positive *social influence* (coordination) in their neighbourhood. A second key difference is that the necessary proportion of neighbours in the q-linked set for a player to be in the q-linked set depends on her *degree*. In the example introduced on Figure 12, player 5 would require at least 47% of her neighbours in the 1.6-linked set to be in it. This proportion rises to 56% for players 7 and 8, while players 1 to 6, 9 and 12 simply don't have enough neighbours to be in the 1.6-linked set.

Our next result summarises the comparative statics at the ME.

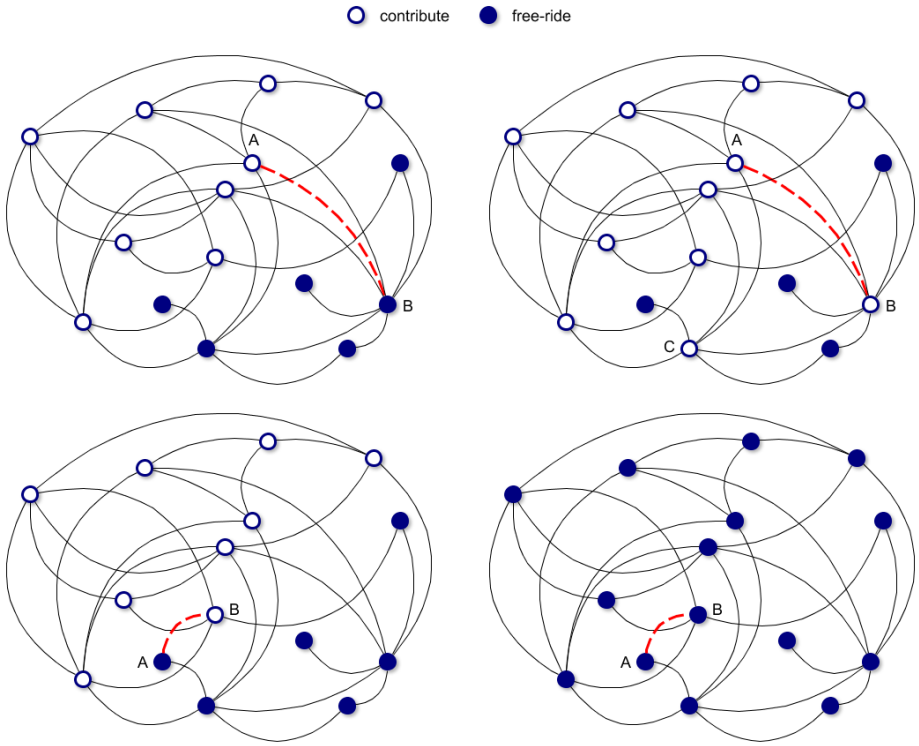
Proposition 2.1 *Suppose that Assumption 2.1 holds. At the unique ME, total contributions are weakly decreasing in γ. For fixed γ and n, the maximal contribution level:*

1. *Weakly decreases (stays the same) with the deletion (addition) of a link between two players in the γ-linked set;*
2. *Weakly increases with the addition or the deletion of a link between two players outside the γ-linked set;*
3. *May increase or decrease (stays the same) with the addition (deletion) of a link between a player in the γ-linked set and a player outside the γ-linked set.*

Note first that a decreasing γ always increases players' incentives to contribute as it decreases the cost of contribution. Decreasing γ makes the requirement on players' minimal neighbourhood influence less stringent. This weakly increases the set $\mathcal{Q}_\gamma$, therefore increasing total contributions at equilibrium.

The effects of network structure on the ME are more subtle. First, the effect on total contributions of adding (or removing) a link depends on its effect on the γ-linked set. In particular, adding a link between two players already in the γ-linked set will leave it unchanged, while removing a link between two such players can only reduce it. Second, adding or removing a link between two players outside of the γ-linked set can never reduce it, as it does not impact the payoffs of any player in the γ-linked set. However, it can increase the γ-linked set: in particular, adding a link between two players outside the γ-linked set may result in both players joining it, while removing a link between a materialist and a conditional cooperator, for example, can induce the latter to switch to contribution. Lastly, adding a link between a player in the γ-linked set and another one outside the γ-linked set can increase, decrease, or have no effect on total contributions. To see why, consider a contributing conditional cooperator i and a free-riding conditional cooperator j. If, in the initial ME, i has only a small net incentive to contribute whereas j has a large net incentive to free-ride, then linking the players will lead i to switch to $x_i = 0$. Conversely, if the magnitude of the incentives happens to be the other way round, then the reverse will hold, and if both players have large incentives for their respective initial actions, then adding a link will make no difference. Figure

Figure 13: The effect of adding a link on the γ-linked set



Top panel: Adding a link between players A and B increases total contributions at the ME. **Bottom panel**: Adding a link between players A and B drives total contributions to 0 at the ME.

13 shows that adding one link to the graph of Figure 11, even when all players are conditional cooperators, can have either positive or negative impact on the ME.

2.4 Susceptibility and influence

We now turn our attention to players' *influence*. In particular, in a given network, who are the most 'influential' players? How does a player's influence depend on how 'central' she is?

Consider a network G, and fix $\gamma \in \mathbb{R}_+$ and $\theta \in \Theta$. Denote the ME by $\mathbf{x}^*$. Take a player i with $\theta_i = \theta_R$, and switch her type to θ_M.[49] Denote the new ME by $\mathbf{x}'$, with $x'_i = 0$ and $x'_j \leq x^*_j$ for all $j \in N$. We study how and whether i's switch to a materialist type induces other players

[49]The opposite case, i.e. where $\theta_i = \theta_M$ initially, is analogous and is thus omitted.

to switch to free-riding, and how this influence depends on the network structure.

First, a player j is *susceptible* to player i, or *i-susceptible,* if and only if $x_j^* \neq x_j'$; thus j is i-susceptible if and only if j switches to free-riding following i's switch to θ_M. Second, we define i's *influence*, denoted by $\mathcal{I}_i(\cdot)$, as the proportion of i-susceptible players in $N \setminus \{i\}$:

$$\begin{aligned} \mathcal{I}_i(\theta, \gamma, G) &= \frac{\sum_{j \neq i} \left(x_j^* - x_j'\right)}{n-1} \\ &= \frac{|\mathcal{S}^i(G)|}{n-1} \end{aligned} \tag{2.7}$$

where $\mathcal{S}^i(G)$ denotes the set of i-susceptible players (excluding i) at the unique ME $\mathbf{x}^*$.[50] Lastly, denote by $r_j^*(\cdot)$ the minimal number of j's neighbours switching to free-riding necessary for j to switch to free-riding at the unique ME:

$$r_j^*(\theta, \gamma, G) := \min_{r \in \mathbb{N}_0} \left\{ r : \Psi\left((c_j^* - r,), (d_j^* + r)\right) \leq \gamma \right\} \tag{2.8}$$

If $r_j^* = 0$, then j does not need any more neighbours switching to free-riding to prefer free-riding herself. This means that j is already free-riding, i.e. $x_j^* = x_j' = 0$.

Remark 2.1 *Suppose that Assumption 2.1 holds, and fix γ and G. Then any player i has positive influence if and only if $x_i^* = 1$ and there exists some player $j \in N_i(G)$ such that $r_j^* = 1$.*

Remark 2.1 states two jointly necessary and sufficient conditions for player i's influence to be greater than 0. *First*, i must be i-susceptible herself. Otherwise, her type switch does not affect her action (as $x_i^* = x_i' = 0$), which trivially leaves the equilibrium unchanged. *Second*, i must have at least one "unconditional follower" in her neighbourhood, i.e. a player j who only needs i to switch her action to switch hers. If i does not have such a neighbour, then even if she switches her action, none of her neighbours (and, thus, none of her neighbours' neighbours, and so on) will switch theirs. Hence i's switch will not spread.

[50] Note that if $x_i^* \neq x_i'$, we say that i is herself i-susceptible. Note also that materialist players, by definition, are never susceptible to other players as they always free-ride at any equilibrium.

Proposition 2.2 *Suppose that Assumption 2.1 holds, and fix γ and G. Consider a player i with positive influence. Then, a player $j \neq i$ is i-susceptible if and only if $r_j^* > 0$ and j is connected to at least r_j^* other i-susceptible players.*

Corollary 2.1 *The set of players $\mathcal{S}^i(G)$ contains no non-empty subset $A \subseteq \mathcal{S}^i(G)$ such that every $j \in A$ has strictly fewer than r_j^* links with players in $\mathcal{S}^i(G) \setminus A$.*

Together, Proposition 2.2 and Corollary 2.1 offer an important result: essentially, $\mathcal{S}^i(G)$ *contains players who are sufficiently interconnected, but does not contain any overly- interconnected sub-group.* In other words, a player i is influential if she is connected to a large set of players whose interconnection will be enough to allow her influence to spread (Proposition 2.2), but not so much as to allow players of a subgroup of players to mutually 'immunise' each other (Corollary 2.1). Put metaphorically, a player in a network is influential if she can '*divide just enough to conquer*'.

The subtlety of these results warrants more detailed discussion. Proposition 2.2 tells us that to be i-susceptible, player j must be connected to enough other i-susceptible players.[51] If j is connected to too few i-susceptible players (i.e. fewer than r_j^*), then, from (2.8), j will not be influenced enough to prefer to switch her own action. Corollary 2.1 tells us that i-susceptible players must not be too interconnected; the following reasoning explains why. Suppose *a contrario* that there exists a subset $A \subseteq \mathcal{S}^i(G)$ such that for all $j \in A$, the number of links that j has with players in $\mathcal{S}^i(G) \setminus A$ is smaller than r_j^*. Observe that if all players in A contribute, it follows from definition (2.8) that they then prefer to contribute, which entails that they must contribute at the new ME. Therefore, while *some* interconnection is necessary for influence, *too much* interconnection kills it.

[51] Note that $\mathcal{S}^i(G)$, like the q-core introduced in Section 2.2, is defined reflexively. A major difference, however, is that $\mathcal{S}^i(G)$ is defined with respect to a single player, i, who is always in $\mathcal{S}^i(G)$ whenever the set is non-empty. Also, while the q-core is a maximal set, $\mathcal{S}^i(G)$ is "minimal" in the sense of Corollary 2.1.

Proposition 2.3 *Suppose that Assumption 2.1 holds, and fix γ and G. Then, for any $i \in N$, i's influence:*

1. *Weakly decreases with the addition or deletion of a link between two i-susceptible players;*
2. *Weakly increases with the addition or the deletion of a link between two non-i-susceptible players if the ME is left unchanged, and may go either way otherwise;*
3. *May increase, decrease or stay the same with the addition or the deletion of a link between an i-susceptible player and a non-i-susceptible player.*

Proposition 2.3 explores how changes to the network affect a player i's influence. Proposition 2.3(1), first, states a robust result: adding or removing a link in the set of i-susceptible players always reduces i's influence. Adding a link *increases* the interconnection between i-susceptible players, which Corollary 2.1 tells us can reduce i's influence. Conversely, removing a link between two i-susceptible players *reduces* the interconnection between i-susceptible players, which we know from Proposition 2.2 can also reduce i's influence. Second, note that Proposition 2.3(2) stems from Corollary 2.1: fixing the ME, removing a link between a pair of non-i-susceptible players may decrease their interconnection enough for i's influence to spread (e.g. by turning one or both of them into being i-susceptible). Likewise, adding a link between two non-i-susceptible players, e.g. one contributing player i and one materialist j, may make the latter i-susceptible, therefore increasing i's influence. Proposition 2.1, however, tells us that adding or removing a link can change the maximal equilibrium: in such case, i's influence may ultimately increase or decrease.[52] Finally, Proposition 2.3(3) states that adding or removing links between i-susceptible players and non-i-susceptible players has an ambiguous effect on i's influence.

We conclude with an observation on how a player's influence relates to her *centrality* (e.g. degree, eigenvector, betweenness centrality). A first intuitive guess would be that the more central a player is, the more influential she must be. This intuition, however, is incorrect.

[52]For example, adding a link between a contributing conditional cooperator k and a materialist player j may induce k to switch to free-riding. This can reduce i's influence by inducing in turn an i-susceptible player to switch to free-riding (recall that already free-riding players cannot be i-susceptible). Conversely, k's switch can also decrease r_l^* for some player l who may then become i-susceptible, which increases i's influence.

Table 1: Figure 14 and nodes' centrality and influence

Players	**Degree**	**Eigenvector**	**Betweenness**	$\mathcal{I}_i$ $(\gamma = 0.4)$	$\mathcal{I}_i$ $(\gamma = 1.5)$
1	0.71	0.47	0.63	**1.00**	**1.00**
2 & 3	0.29	0.14	0.06	0.14	**1.00**
4 & 5	0.29	0.12	0.02	0.14	**1.00**
6 & 7	0.43	0.36	0.25	**1.00**	**1.00**
8	**1.00**	**1.00**	**1.00**	0.00	**1.00**
9 to 12	0.57	0.77	0	0.00	0.00

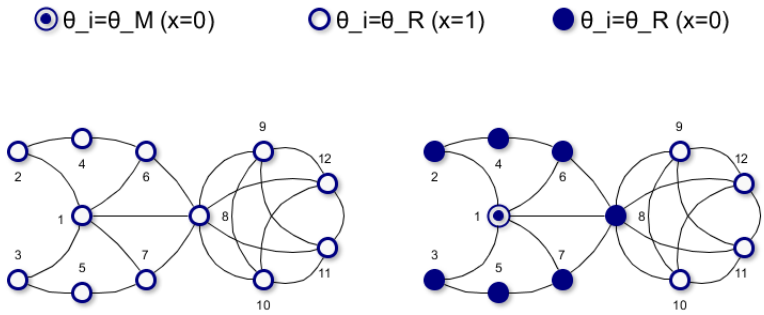
Players' centrality/influence indices expressed in proportion to the highest centrality/influence among nodes.

Remark 2.2 *In general, a player's influence and her centrality do not necessarily coincide.*

Remark 2.2 rests on two key ideas. *First*, note that for a given $\mathbf{r}^* = \left(r_1^*, ...r_j^*, ..r_n^*\right)$, the vector of the minimal numbers of additional free-riding neighbours necessary for players to switch to free-riding at the ME, $\mathcal{S}^i(G)$ is solely determined by G, as in the case of centrality metrics. However, $\mathbf{r}^*$ also depends on γ and θ. Hence, for any i, the set of i-susceptible players, and thus i's influence, depend on G, γ and θ. *Second*, our concept of influence captures something novel: a player is influential if she is connected to players who are interconnected, but not too much. This contrasts with, for instance, eigenvector centrality, which rests on the idea that central players are those who are well-connected to other central players.

Table 1 explores this distinction by presenting the degree, eigenvector, and betweenness centralities of players in Figure 14, as well as their influence for different values of γ. Suppose that preferences are given by (2.4) with $\alpha = 1$ and $\beta = 1.5$. A first remark is that player 8 is, by all measures, the most central player; nevertheless, she has no influence when $\gamma = 0.4$. Conversely, players 6 and 7 have relatively low degree, eigenvector and betweenness centrality, but have the highest influence (for both $\gamma = 0.4$ and $\gamma = 1.5$). Note also that the increase in γ from 0.4 to 1.5 pushes the influence of all players up: this is because when the cost of contributing is higher, contributing players require fewer free-riding neighbours to prefer to free-ride. In other words, it takes less to convince them to switch.

Figure 14: Influence



Left: initial graph, ME with $\theta_i = \theta_R$ for all $i \in N$, and $\gamma = 0.4$. **Right**: ME with θ_1 switched to θ_M.

2.5 Network design

In this section, we turn to the problem faced by a planner who has to design the network and wishes to maximise the expected contribution level for some fixed number of players. We seek to answer the following question: What network structure maximises expected contributions? In particular, are denser structures more resilient to 'bad apples' (i.e. materialists) than sparser ones? Are disconnected networks more robust to free-riding than connected ones? How does the optimal network structure change with the incidence of materialist players?

Suppose that players' types are i.i.d., and the probability of any player being a materialist is given by $p \in (0,1)$ (so the probability of being a conditional cooperator is $1-p$). The designer's decision boils down to choosing G so as to maximise the *ex ante* probability that any player $i \in N$ cooperates given G. This problem is symmetric for all $i \in N$, and we thus restrict attention to regular networks.[53] Formally, the designer's problem is to choose $G(N)$ so as to:

$$\max \sum_{i \in N} \mathbb{E}_{p,\tilde{\theta}} \left(x_i^* \left(\tilde{\theta}, \gamma, G(N) \right) \right) \tag{2.9}$$

where $x_i^*(\cdot)$ is i's action at the unique ME $\mathbf{x}^*$ for the chosen network $G(N)$, given γ and

[53] For simplicity, we assume that the designer's decision yields no remainder.

θ, and expectations are taken over type profiles $\tilde{\theta}$. Lastly, define a *clique of degree* k as a set of $k+1$ players all connected to one another. An *isolated clique* is a clique of players who have no links to players outside the clique. We define a *network of cliques of degree* k as a network comprising only isolated cliques of degree k.

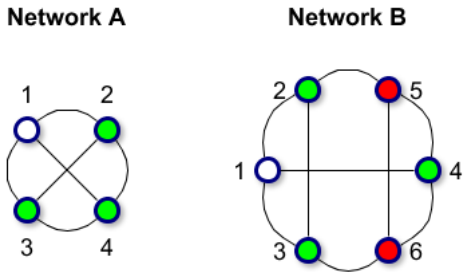
Theorem 2.2 *Suppose that Assumption 2.1 holds and that players' types are i.i.d., with* $p \equiv Pr(\theta_i = \theta_M) \in (0,1)$ *for all* $i \in N$. *For any* $\gamma > 0$ *and generic* p, *there exists a unique integer* $k^*(p) \in \{\underline{k}, \ldots, n-1\}$ *such that the network of cliques of degree* $k^*(p)$ *maximises* ex ante *total contributions, where* $\underline{k}$ *is the smallest integer such that* $\Psi(1, \underline{k}, 0|\theta_R) \geq \gamma$.

Theorem 2.2 underscores two crucial elements of the designer's decision. First, any player must be minimally connected: this stems from the fact that any player with fewer than $\underline{k}$ connections to other players is sure to free-ride. Hence, the designer can always do better *ex ante* by raising all players' degree to at least $\underline{k}$.

Second, the network that maximises *ex ante* total contributions is always the network of cliques of degree $k^*(p)$, with $k^*(p) \in \{\underline{k}, \ldots, n-1\}$. The reason is that the network of cliques of degree $k^*(p)$ is the network that both maximises the chances of clustering of conditional cooperators (enabling them to contribute at equilibrium) while minimising materialists' expected influence. The intuition underlying the proof is as follows: in a clique of degree k, the probability that any conditional cooperator i cooperates depends solely on the probability that enough other players in the clique (i.e. in i's k connections) are also conditional cooperators. This is due to the clique's maximal *cohesion*. In any other network, that probability must be adjusted for (and hence reduced by) the probability that those k players are also connected to enough cooperating players, since players' neighbourhoods do not necessarily coincide. The probability that any player i cooperates in this case is thus always lower than in the clique of degree k as free-riding can easily spread, while its effect is contained in the clique. Figure 15 illustrates this argument, assuming that player 1 is a conditional cooperator and that $\underline{k} = 3$.

We next ask how the size of the cliques in the the unique optimal network of cliques of degree $k^*(p)$ changes with p. To fix ideas, note that the designer faces a tradeoff when

Figure 15: Isolated cliques vs regular networks



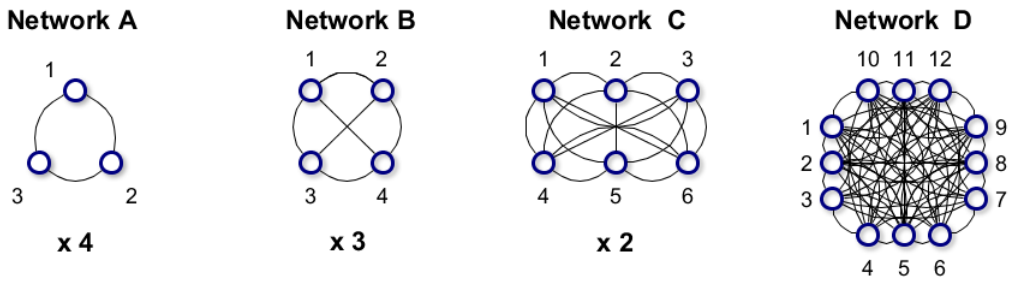
Network A: A clique with $k_i = 3$ for all $i \in N$. Player 1, assumed to be a conditional cooperator, cooperates if and only if players 2, 3 and 4 are conditional cooperators. The probability of this event is $(1-p)^3$. **Network B**: A regular network with $k_i = 3$ for all $i \in N$. Player 1 cooperates if and only if not only players 2, 3 and 4, but also 5 and 6 are conditional cooperators. The probability of this event is $(1-p)^5$.

choosing k^*. On the one hand, an increasing k^* entails the benefit of increasing the maximal number of materialists allowed in a clique for cooperation not to break down. On the other hand, an increasing k^* also comes with the cost attached to an increasing number of "draws" and, thus, expected number of materialists in the clique.

Proposition 2.4 *Suppose that Assumption 2.1 holds and that players' types are i.i.d., with* $p \equiv Pr(\theta_i = \theta_M) \in (0,1)$ *for all* $i \in N$. *For any* $\gamma > 0$ *and* p, *the size of the cliques in the unique optimal network of cliques of degree* $k^*(p)$ *weakly decreases with* p.

Proposition 2.4 presents a subtle result. Indeed, observe that increasing p yields two opposing effects on $k^*(p)$. First, a higher p raises the expected incidence of materialists. This effect pushes $k^*(p)$ *down*, as the designer wants to reduce the expected number of materialists in any conditional cooperator's neighbourhood. However, a higher p also reduces the expected incidence of conditional cooperators. This pushes $k^*(p)$ *up,* as the designer wants to increase the expected number of conditional cooperators in any conditional cooperator's neighbourhood. The former effect, it turns out, always dominate the former. To provide intuition, observe that when p is low, materialists are unlikely and their effect can likely be more than offset by the presence of conditional cooperators. Hence, the optimal cliques are large. Conversely, when p is high, the designer knows that a large clique will probably result in

Figure 16: Network Design



a "spoiled barrel" due to a large (expected) number of materialists. In such case, the designer prefers smaller cliques: while each of those cliques is likely to be "spoiled", there is a chance that some will not, resulting in higher cooperation (on average) than in any other network.

We proceed with an example to illustrate our results. Suppose that a designer needs to choose $G(N)$ for $N = 12$, and let $\gamma = 1.1$. Theorem 2.2 tells us that the designer's candidate options are given by the networks A, B, C and D on Figure 16. Suppose that conditional cooperators' social payoffs are given by (2.4) with $\alpha = 1.25$ and $\beta = 1$. It can first be shown that B always dominates A. The reason is that the designer knows that cooperation in any clique in A breaks down at equilibrium with any materialist, while any clique in B admits (at most) one materialist. The tradeoff between increased number of draws and increased allowed number of materialists is thus trivial: with probability p, the additional player is a materialist, and the likelihood of cooperation in the clique of size 4 is the same as in the clique of size 3. With probability $1 - p$, the additional player is a conditional cooperator, and the likelihood of cooperation in the clique of size 4 is then strictly higher than in the clique of size 3. Hence, the designer can always raise his expected payoffs by increasing the cliques' size from 3 to 4. Whether the designer prefers B, C or D however depends on p, and it can be easily computed that the designer prefers B for any $p \in [0.5, 1]$ and D for any $p \in [0, 0.5]$.

2.6 Discussion: social influence at the workplace

In this section, we discuss our results in the context of an important application: work morale and peer influence at the workplace. We show that our model brings novel insights to many questions related to co-workers' interactions and their impact for team performance. In particular, how do co-workers influence each other with respect to productive effort within firms? When can a 'bad apple' spoil team spirit and productivity in a team? How can one best structure a team to best promote desirable behaviour and prevent contagion of shirking?

We map a typical workplace situation onto our model as follows. Nodes are co-workers. In a firm, any given worker is likely to interact only with a subset of all other workers: links represent work relationships. Each worker decides whether to exert effort ($x = 1$) or not ($x = 0$). We suppose that the manager cannot observe employees' effort directly, while workers do observe the effort of the co-workers they are linked to. Lastly, we assume that exerting effort creates material positive externalities on all other workers: for example, efforts increase the team or firm's profits or likelihood of success, from which all workers benefit.

2.6.1 Reciprocity, work morale and networks

Evidence unarguably demonstrates the existence of reciprocal preferences at the workplace. Productive or cooperative workers increase their co-workers' motivation to work hard (or make them feel guilty when they do not exert effort), while shirkers create a feeling of inequity:

> ... [in the presence of a] withholder of effort, teammates are unlikely to be motivated to contribute to the collective pool of ideas... perception of inequity will arise when group members compare their own contributions to those of a withholder of effort in their team, and will result in a desire to restore equity by reducing contributions. (Felps et al., 2006: 191-203)

Our first main finding, Theorem 2.1, yields a key insight with respect to reciprocity and effort at the workplace: to exert effort, a worker must be connected enough, and in high enough proportion, to co-workers who themselves exert effort. This prediction finds support in the empirical literature. Hesselius et al. (2009) exploit a large natural random experiment and

report clear evidence that workers display a "reciprocal type of preferences and/or display fairness concerns" (p. 585). They show that "observing a sudden increased absence level" in a worker's network of co-workers "may induce resentment and lead to ill feelings towards the shirking co-workers" (Hesselius et al., 2009: 585-91), which in turn further fosters shirking. Ichino and Maggi (2000) argue that an agent's shirking level increases with the shirking level of her co-workers' in her network. In contrast, when a worker "is surrounded by a group that works very hard, shirking may induce... a sharper feeling of guilt" (Ichino and Maggi, 2000: 1066). Bandiera et al. (2010) and Mas and Moretti (2009), similarly, find that more (less) productive workers increase (decrease) the productivity of their peers in their network.[54]

2.6.2 'Bad apples' and workers' influence

Our framework brings novel insights to the study of workers' influence, especially 'bad apples', on their peers. Our analysis, summarised in Proposition 2.2 and Corollary 2.1, offers the following prediction: a player will be influential if she is connected to players who are sufficiently interconnected, but not too much. This criterion is likely to be satisfied in small, dense and interdependent teams, where all workers are connected to one another, but not to enough workers elsewhere to be unsusceptible to a single bad apple. This prediction finds support in the empirical literature: "[a bad apple's] destructive behaviour [is] particularly impactful in *small* groups... in *interdependent* teams where people depend on each other, [...] intense psychological reactions [to bad apples] are more likely to spill over" (Felps et al., 2006: 180-190; italics added). Our analysis also suggests that workers with low degree (and therefore low r_j^*) will be particularly susceptible to bad apples. Felps et al. (2006) and Brass et al. (1998) review evidence supporting this result. They propose that a worker's susceptibility to a bad apple depends on the ratio of contacts said worker has with the bad apple compared to those

[54]Theorem 2.1 yields another insight: due to network effects, a worker's effort decision may depend on the impact of distant workers in the network. Experimental evidence strongly suggests the existence of such a contagion effect for both free-riding (Gracia-Lazaro et al., 2012; Suri and Watts, 2011; and Grujic et al., 2010) and cooperation (Rand et al., 2014; Jordan et al., 2013).

she has with other co-workers.[55]

Lastly, evidence suggests that workers often take action to reduce a bad apple's "degree" or influence on their group. A typical response to shirkers is rejection, aiming at "reducing social interaction... [or to] restructure work to decrease task interdependence" with those who withhold effort (Felps et al., 2006:186; see also Lepine et al., 1997). Likewise, "companies fire shirkers... to re-establish the work morale of the rest... [M]otivated workers may prefer that bad apples are fired because they do not like being suckered by their colleagues and because it re-establishes beliefs about others' team-spirit." (Gächter, 2006: 22). This is in line with our analysis.

2.6.3 Optimal team design

Our model provides several novel insights with respect to optimal team design, summarised in Theorem 2.2 and Proposition 2.4. A first result is that teams with a certain level of interdependence (i.e. minimal degree) are always preferable to isolated workers. The available empirical evidence is compelling: Falk and Ichino (2006) find clear evidence that on average, workers are less productive when isolated than when in teams. They explain that "people working in groups feel some pressure to keep up with the efforts of those around them, and/or the most productive workers pressure others into working harder" (Falk and Ichino, 2006: 40), which is consistent with our result.

Second, our analysis suggests that when 'bad apples' are numerous, interdependence (i.e. density) in a large network of workers comes with higher risk of contagion. Experiments of PGGs on complete networks demonstrate that when the proportion of "unconditional free-riders" (i.e. bad apples) in the population is significant, the global contribution level of the population is driven to zero due to the negative influence free-riders exert on conditional cooperators (see e.g. Fehr and Gächter, 2000b, for review). In contrast, Rand et al. (2014)

[55]This prediction is supported in other contexts, e.g. crime and teenage delinquency. For example, Rees and Pogarsky (2011) find that teenagers, when faced with bad behaviour by their best friends, are less likely to behave badly themselves the greater the number of other social contacts they have (see also Akers and Jensen, 2006). This is consistent with our analysis.

show that cooperation can be maintained in sparser networks. They explain that "when interactions are structured, such that people only interact with their network neighbours rather than the whole population, the emergence of clustering is facilitated" (Rand et al., 2014: 17093). In other words, reducing the interdependence of workers imposes a limit on how influential a bad apple can be in a team or network. This can result in greater cooperation, as our results show.[56]

Conversely, our analysis predicts that when the risk of bad apples is low, limited interdependence (e.g. in small isolated cliques) might make workers unnecessarily susceptible to potential bad apples. For example, in a small clique, a single bad apple may be sufficient to drive cooperation to zero, as discussed above. In contrast, experimental evidence suggests that groups consisting only of conditional cooperators are characterised by high and sustained cooperation (Gunnthorsdottir et al., 2007; Gächter and Thöni, 2005). Grouping workers in large interdependent units is thus optimal when shirkers are uncommon.

2.7 Conclusion

In Chapter 2, we undertook the study of conditional cooperation on social networks. We characterised the (unique) maximal Nash equilibrium (ME) of a one-shot public good game for any fixed network, assuming only weak conditions on conditional cooperators' utility functions that capture the defining features of reciprocity. At the ME, a novel measure of network structure, the *q-linked set*, fully determines the set of players who contribute. The q-linked set consists in the largest set of conditional cooperators with both enough *cohesion* and *density*. We gave a novel characterisation of a player's *influence* on others, and showed that while the set of players whom i influences must be interconnected to some extent, it must not be *too* interconnected. We then studied the decision of a manager who designs the network to

[56]Note that our result on small isolated cliques when p is high relates strongly to the optimal architecture of terrorist or revolutionary organisations, i.e. in isolated "cells". Cells' main advantage is the robustness to outside threats (e.g. police) they bring to the whole network, as detection does not spread to other cells (see Baccara and Bar-Isaac, 2008). The tradeoff we have is similar, but we highlight the role of *internal* threats: when some members shirk, they can entail the breakdown of the whole organisation.

maximise expected contributions under an uncertain type profile. We showed that the *ex ante* optimal network is formed by isolated cliques of degree k^*, with k^* decreasing with the probability of any player being a materialist. Lastly, we discussed our results in the context of one important application: work morale and peer influence at the workplace. Our results yield testable predictions that accord well with the available evidence.

2.8 References

1. Ali, S. N. and D. A. Miller, 2013. Enforcing Cooperation in Networked Societies. *Mimeo*, University of California, San Diego.

2. Ali, S. N. and D. A. Miller, 2016. Ostracism and Forgiveness. *American Economic Review*, vol. 106(8), pp. 2329-48.

3. Akers, R. L. and G. F. Jensen, 2006. The Empirical Status of Social Learning Theory of Crime and Deviance: The Past, Present and Future. In F. T. Cullen, J. P. Wright and K. R. Blevins (eds). *Taking Stock: The Status of Crimino-Logical Theory.* Transaction Publishers: New Brunswick.

4. Baccara, M. and H. Bar-Isaac, 2008. How to Organize Crime? *Review of Economic Studies*, vol. 75, pp. 1039-67.

5. Bandiera, O., Barankay, I. and I. Rasul, 2010. Social Incentives in the Workplace. *Review of Economic Studies*, vol. 77, pp. 417-58.

6. Bollobás, B., 1984. The evolution of sparse graphs. Graph theory and Combinatorics (Cambridge, 1983). Academic Press, London, pp. 35–57.

7. Bolton, G.E. and A. Ockenfels, 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, vol. 90(1), pp. 166-93.

8. Bramoullé, Y. and R. Kranton, 2015. Games Played on Networks, in Y. Bramoullé, A. Galeotti and B. Rogers (eds), *Oxford Handbook on Economics of Networks,* Oxford: Oxford University Press.

9. Brass, D. J., Butterfield, K. D. and B. C. Skaggs, 1998. Relationships and Unethical Behavior: A Social Network Perspective. *The Academy of Management Review*, vol. 23(1), pp. 14-31.

10. Burt, R., 1992. *Structural Holes: The social structure of competition*, Cambridge (Massachusetts): Harvard University Press.

11. Cabrales, A., Gottardi, P. and F. Vega-Redondo, 2016. Risk-Sharing and Contagion in Networks. *Mimeo,* University College London.

12. Carpenter, J., Kariv, S. and A. Schotter, 2012. Network Architecture, Cooperation and Punishment in Public Good Experiments. *Review of Economic Design,* vol. 16(2-3), pp. 93-118.

13. Cassar, A., 2007. Coordination and Cooperation in Local, Random and Small World Networks: Experimental Evidence. *Games and Economic Behavior*, vol. 58(2), pp. 209-30.

14. Charness, G. and M. Rabin, 2002. Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics*, vol.117, pp. 817-69.

15. Chaudhuri, A., 2011. Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature. *Experimental Economics*, vol. 14, pp. 47-83.

16. Coleman, J.S, 1990. *Foundations of Social Theory*, Cambridge (Massachusetts): Harvard University Press.

17. Dufwenberg, M. and G. Kirchsteiger, 2004. A Theory of Sequential Reciprocity. *Games and Economic Behavior*, 47(2), 268-98.

18. Ellison, G., 1994. Cooperation in the Prisoner's Dilemma with Anonymous Random Matching. *Review of Economic Studies*, 61, 567-88.

19. Erol, S. and R. Vohra, 2014. Network Formation and Systemic Risk. *PIER Working Paper No. 15-001.*

20. Falk, A. and A. Ichino, 2006. Clean Evidence on Peer Effects. *Journal of Labor Economics*, vol. 24(1), pp. 39-57.

21. Falk, A., Fischbacher, U. and S. Gächter, 2013. Living in Two Neighborhoods – Social Interaction in the Lab. *Economic Inquiry*, vol. 51(1), pp. 563-78.

22. Fehr, E. and K. Schmidt, 1999. A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics*, vol. 114, pp. 817-68.

23. Fehr, E. and S. Gächter, 2000a. Cooperation and punishment in public goods experiments. *American Economic Review*, vol. 90(4), pp. 980–94.

24. Fehr, E. and S. Gächter, 2000b. Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, vol. 14(3), pp. 159-81.

25. Fehr, E. and S. Gächter, 2002. Altruistic Punishment in Humans. *Nature*, vol. 415, pp. 137-40.

26. Felps, W., Mitchell, T. R. and E. Byington, 2006. How, When, and Why Bad Apples Spoil the Barrel: Negative Group Members and Dysfunctional Groups. *Research in Organizational Behavior*, vol. 27, pp. 175-222.

27. Fowler, J. H. and N. A. Christakis, 2010. Cooperative Behaviour Cascades in Human Social Networks. *PNAS*, vol. 107(12), pp. 5334-38.

28. Fu, F., Hauert, C., Nowak, M. A. and L. Wang, 2008. Reputation-Based Partner Choice Promotes Cooperation in Social Networks. *Physical Review E, Statistical, Nonlinear and Soft Matter Physics*, 78(2pt2), 026117.

29. Gächter, S. and C. Thöni, 2005. Social Learning and Voluntary Cooperation Among Like-Minded People. *Journal of the European Economic Association*, vol. 3(2-3), pp. 303-14.

30. Gächter, S., 2006. Conditional Cooperation: Behavioral Regularities from the Lab and the Field and their Policy Implications. Discussion paper, University of Nottingham.

31. Gallo, E. and C. Yan, 2015. The Effects of Reputational and Social Knowledge on Cooperation. *PNAS*, vol. 112(2), pp. 3647-52.

32. Granovetter, M., 2005. The Impact of Social Structure on Economic Outcomes. *Journal of Economic Perspectives*, vol. 19(1), pp. 33-50.

33. Gouldner, A. W., 1960. The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review*, vol. 25(2), pp. 161-78.

34. Gracia-Lazaro, C., Ferrer, A., Ruiz, G., Tarancon, A., Cuesta, J. A., Sanchez, A. and Y. Moreno, 2012. Heterogeneous Networks Do Not Promote Cooperation when Humans Play a Prisoner's Dilemma. *PNAS*, vol. 109(32), pp. 12922-26.

35. Grujic, J., Fosco, C., Araujo, L., Cuesta, J.A. and A. Sanchez, 2010. Social Experiments in the Mesoscale : Humans Playing a Spatial Prisoner's Dilemma. *PLoS ONE*, 5(11), e13749.

36. Gunnthorsdottir, A., Houser, D. and K. McCabe, 2007. Disposition, History and Contributions in Public Goods Experiments. *Journal of Economic Behaviour and Organization*, vol. 62(2), pp. 304-15.

37. Haag, M. and R. Lagunoff, 2006. Social Norms, Local Interaction, and Neighborhood Planning. *International Economic Review*, vol. 47(1), pp. 265-96.

38. Hesselius, P., Johansson, P. and J. P. Nilsson, 2009. Sick of Your Colleagues' Absence? *Journal of the European Economic Association*, vol. 7(2-3), pp. 583-94.

39. Hopfensitz, A. and E. Reuben, 2009. The Importance of Emotions for the Effectiveness of Social Punishment. *Economic Journal*, vol. 119, pp. 1534-59.

40. Huck, S., Kubler, D. and J. Weibull, 2012. Social Norms and Economic Incentives in Firms. *Journal of Economic Behavior and Organization*, vol. 83, pp. 173-85.

41. Ichino, A. and G. Maggi, 2000. Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm. *Quarterly Journal of Economics*, vol. 115(3), pp.1057-1090.

42. Jackson, M. and Y. Zénou, 2014. Games on Networks. In P. Young and S. Zamir (eds), *Handbook of Game Theory (Vol. 4)*, Elsevier.

43. Jordan, J. J., Rand, D. G., Arbesman, S., Fowler, J. H. and N. A. Christakis, 2013. Contagion of Cooperation in Static and Fluid Social Networks. *PLoS One*, 8(6), e66199.

44. Kandel, E. and E. Lazear, 1992. Peer Pressure and Partnerships. *Journal of Political Economy*, vol. 100, pp. 801-13.

45. Kandori, M., 1992. Social Norms and Community Enforcement. *Review of Economic Studies*, vol. 59(1), pp. 63-80.

46. Lepine, J. A., Hollenbeck, J. R., Ilgen, D. R. and J. Hedlund, 1997. Effects of Individual Differences on the Performance of Hierarchical Decision-Making Teams: Much More than g. *Journal of Applied Psychology*, vol. 82(5), pp. 803-11.

47. Mas, A. and E. Moretti, 2009. Peers at Work. *American Economic Review*, vol. 99(1), pp. 112-45.

48. Mauss, M., 1954 [1950]. The Gift. Glencoe: Free Press.

49. Morris, S., 2000. Contagion. *Review of Economic Studies*, vol. 67(1), pp. 57-78.

50. Milgrom, P. and P. Roberts, 1990. Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities. *Econometrica*, vol. 58(6), pp. 1255–77.

51. Nava, F., 2015. Repeated Games on Networks, in Y. Bramoullé, A. Galeotti and B. Rogers (eds), *Oxford Handbook on Economics of Networks*, Oxford: Oxford University Press.

52. Nowak, M., 2012. Evolving Cooperation. *Journal of Theoretical Biology*, vol. 299, pp. 1-8.

53. Rabin, M., 1993. Incorporating Fairness into Game Theory and Economics. *American Economic Review*, vol. 83(5), pp. 1281-301.

54. Rand, D. G., Arbesman, S., and N. A. Christakis, 2011. Dynamic Social Networks Promote Cooperation in Experiments with Humans. *PNAS*, vol. 108(48), pp. 19193-98.

55. Rand, D. G., Nowak, M. A., Fowler, J. H. and N. A. Christakis, 2014. Static Network Structure Can Stabilize Human Cooperation. *PNAS*, vol. 111(48), pp. 17043-98.

56. Rees, C. and G. Pogarsky, 2011. One Bad Apple May Not Spoil the Whole Bunch: Best Friends and Adolescent Delinquency. *Journal of Quantitative Criminology*, vol. 27, pp. 197-223.

57. Rotemberg, J., 2006. Altruism, Reciprocity and Cooperation in the Workplace. In S.-C. Kolm and J.M. Ythier (eds), *Handbook of the Economics of Giving, Altruism and Reciprocity (Vol 2)*. North-Holland.

58. Simmel, G. 1950. *The sociology of Georg Simmel*, ed. and trans., K. H.Wolff, Glencoe: Free Press.

59. Sobel, J., 2005. Interdependent Preferences and Reciprocity, *Journal of Economic Literature*, vol. 43, pp. 392-436.

60. Suri, S. and D. J. Watts, 2011. Cooperation and Contagion in Web-Based, Networked Public Good Experiments. *PLoS One*, 6(3), e16836.

61. Topkis, D. M., 1979. Equilibrium Points in Nonzero-Sum N-Person Submodular Games. *SIAM Journal of Control and Optimization*, vol. 17(6), pp. 773–87.

62. Vives, X. 2005. Complementarities and Games: New Developments. *Journal of Economic Literature*, XLIII, pp. 437–479.

63. Wolitzky, A. 2013. Cooperation with Network Monitoring. *Review of Economic Studies*, vol. 80, pp. 395-427.

Chapter 3 Ideological games

Alexander Harris[57]

3.1 Introduction

Ideologically-motivated behaviour underlies many social, historical and economic phenomena. In sociology and political science, affiliation to ideological positions can explain outcomes ranging from decisions by justices of the US Supreme Court (Segal and Cover, 1989) to the willingness of local officials to learn from others' mistakes in setting policy on public planning (Butler et al., 2017). A person's ideological stance is not necessarily static: individuals are influenced by those around them, and often try to influence others in turn. The possibility of influencing people's ideological affiliation leads to the notion of "culture wars", in which rival groups within a society attempt to ensure that their own cultural and ideological values prevail among the population (Hunter, 1991). Historical study, meanwhile, highlights the important role played by ideology in diverse major events in the modern era, from the French and American revolutions to foreign policy in the Cold War (Cassels, 1996). In economics, the existence of entire economic systems for the production and allocation of goods, such as state socialism, is explained, at least in part, by ideological beliefs. Yet there is little theoretical work within economics to explain, or even describe, how ideology drives decision-making, how ideologies are sustained and how they can spread at each other's expense. One reason there currently exists little economic theory concerning ideologically-based decision-making may be a long-running perception among researchers that agents influenced by ideology are in some sense "irrational" (Roucek, 1944), in which case they are not the proper object of economic study.

[57]I would like to thank Robert Evans for his extensive comments on this chapter. I also gratefully acknowledge helpful ideas and comments from Matt Elliot, Aytek Erdil, Julien Gagnon, Edoardo Gallo, Sanjeev Goyal, Dan Quigley, Paulina Sliwa and seminar participants in Cambridge. I am grateful for the financial support I received from the Economic and Social Research Council.

In this chapter, I set out a theory of ideology. A necessary starting point is to provide a formal definition. The two main existing approaches in the literature to modelling ideology cast it as a constraint on an agent's choice set (Roemer, 1985) or as a distortion in beliefs (Bénabou, 2008). My approach, in contrast, assumes agents have complete information and unconstrained choice sets, yet nonetheless embeds the notion of ideology in an expected-utility framework, laying the foundations for quantitative models of how ideologies can spread among a population. To do this, I draw on an observation by Sen (1977) regarding a key feature of moral positions: that they require an agent to hold them 'deeply'. In other words, a moral position not only prescribes a ranking over outcomes, but also requires its holder to prefer holding this ranking to holding other rankings. I represent ideology along these lines: in my model, a preference type is a set of first-order preferences over the outcomes of a game Γ, together with a set of ('meta-') preferences over all players' first-order preferences; preference types meeting certain conditions qualify as ideologies. Before playing Γ, there is an initial round of play in which a player chooses whether to invest in changing her opponent's first-order preferences. I refer to the resulting two-stage game as an *ideological game*. Studying ideological games provides a clear account of two reasons why rational players may expend time and effort influencing each other's preference types. The first reason is instrumental: players may be able to achieve better outcomes in Γ if they change their opponent's type. The second is 'ideological': players may simply prefer that their opponents share their first-order preferences.

In section 3.2, I first provide an extended example to motivate my theory of ideology. In this example, I consider an ascetic religion, which stresses the value of rejecting material possessions and consumption wherever possible in one's life. 'Devout' players, who follow the religion, interact with each other and with 'secular' (non-religious) players in their daily lives. These routine interactions are represented by Γ, the "game of life". I use this example to set out in intuitive terms my conception of ideology and to consider the ways in which it determines equilibrium behaviour when different preference types have the opportunity to invest in changing each others' types before playing Γ. In addition to changing others' types,

investment simultaneously ensures a player retains her own type, if enough same-type players also invest.

In section 3.3 I then move to a general setting and provide formal definitions of the concepts introduced in the extended example. I provide a taxonomy of types as being "strongly ideological", "weakly ideological" or "pragmatic" and relate these concepts to the notion of homophily (Proposition 3.1). I show that in each case, the incentive for a player to invest in changing others' types can be decomposed into an "instrumental" component associated with the change in equilibrium outcome of Γ that would result, and an "ideological" component associated with a direct payoff from the change in type profile. A player's incentive to invest to retain her own type can be decomposed along similar lines. A weakly ideological player always prefers to retain her own type, regardless of any instrumental benefits that may result from a better equilibrium outcome in Γ. Strong ideology is a refinement of weak ideology: a strongly ideological player also prefers to change other players' types, regardless of any changes in payoffs in Γ she might receive. A pragmatic player, in contrast, is indifferent between all type profiles and is willing to have her type changed if, according to her new type, she would prefer the resulting equilibrium in Γ to that which would result if her type remained unchanged. I study the investment profiles supported in pure-strategy equilibrium, which in general contain at most two investing types (Proposition 3.2).[58] I define a notion of efficiency according to which an investment profile has positive efficiency if it improves aggregate welfare over the no-invest profile. In any two-player ideological game in which players have different types, if both types are ideological, there is an equilibrium investment profile with negative efficiency (Theorem 3.1), since ideological players finds it optimal to invest to retain their own type. If at least one type is strongly ideological, and costs are low enough, then the negative efficiency equilibrium is the only equilibrium. Finally, I define "perfectly ideological" types as strongly ideological meta-payoffs which, if held by all players, result in the best outcome of Γ as judged by that type. In any pairing of a perfectly ideological player with a pragmatic

[58]This simplifying result is a consequence of the way I choose to specify the "conversion technology", the function that determines final types based on initial types and investment choices.

player, if the cost of investment is low enough to induce the former to invest, there is a unique equilibrium with positive efficiency.

Related literature. Roemer (1985) models revolution as a two-player game through which he suggests two alternative rationalisations of ideological behaviour, where 'ideological"' is taken to be roughly synonymous with 'apparently irrational'. The first is that being known to have an ideology can allow an agent to pre-commit, which is useful in helping achieve his goals; the second is that seemingly extreme actions in the model can in fact be optimal. Bénabou (2008) follows a different approach. Noting that ideologies tend to correlate with important economic and societal outcomes, but are nonetheless often 'immune' to relevant evidence, ideologies in his model are distorted beliefs that are, nonetheless, individually rational, in that agents can gain utility from ignorance. My model also seeks to explain seemingly irrational behaviour in terms of expected-utility maximisation, but, in contrast to Bénabou's approach, always assumes agents have correct beliefs. In so doing, it follows groundbreaking contributions to rational choice theory by Becker and others to show how the expected-utility framework can explain, for example, criminal behaviour (Becker, 1976), reproductive choices (Becker, 1981) and addiction to narcotics (Becker and Murphy, 1988; for empirical support, see Gruber and Koeszegi, 2001). Unlike Becker, however, in my model I need not assume that people are entirely self-regarding, though self-regarding ideologies can be accommodated within the framework presented here, as for example in Appendix C.3.

A key feature of my model is that people can influence each other by altering each other's preferences. In the two-player ideological game of my model, I investigate the hallmarks of players' preferences that determine how much they try to influence their opponents. Alonso and Camara (2016) provide a theoretical model in which a player can design a signal to try to persuade others to take an action she prefers. The authors' description of ideology thus centres on beliefs, whereas my own takes preferences as characterising an ideology and assumes complete information. One key difference between Alonso and Camara's account of ideology and mine is that in my model, agents may want to spread their ideology even if it achieves

nothing instrumentally. For example, a religious enthusiast may try to convert others even though his interactions with them will be unaffected. Theories of persuading others are closely related to models of commitment, in which an agent plays a game against her future self (see e.g. Bénabou and Tirole, 2004; Gul and Pesendorfer, 2001; Dekel, Lipman and Rustichini, 2009).

A possible application of my model is in evolutionary theory. In Appendix C.3, I explore how ideological preferences can be introduced into the indirect evolutionary model of Chapter 1 of this thesis, pointing the way towards future study of the evolutionary origins of ideology. I consider the possibility that ideological (meta-) preferences can be the object of evolutionary selection, in a setting in which agents can invest in persuading others to adopt their ideology. Once a new ideology "mutates", in that an individual forms a meta-preference that she and those around her rank outcomes a certain way, she may find it optimal to persuade others to share this outlook. The approach is novel in allowing preferences to evolve by agents competing directly over what preferences each of them holds. This idea has roots in work by Dawkins (1976), but has not previously been addressed quantitatively, to the best of my knowledge.[59]

A key distinction in this chapter is between first-order and meta-preferences, which is needed in order to model agents as expected utility maximisers when their 'future selves' (i.e. final-stage types) are determined endogenously. In the framework I will present, first-order preferences are defined over the space of lotteries of the set of outcomes in the second stage of a game. Meta-preferences, which players have in the first stage, are then defined over these first-order preferences together with lotteries over final-stage outcomes. I characterise ideological types as sets of meta-preferences in which a player prefers her initial (first-order) preference type to the alternative. However, my definition of meta-payoffs in general allows for a player to prefer that his future self (or the future selves of other players) hold the alternative preference type. Harsanyi (1955) envisages 'ethical' preferences as those which one would choose to hold based on 'social considerations' alone, as opposed to one's 'subjective' preferences (i.e.

[59]Indeed, Dennett (1995) voices scepticism as to the feasibility of such an account.

those preferences that one actually holds). With Harsanyi's notion in mind, we can think of ideological payoffs as being a subset of ethically-motivated (or moral) payoffs. In my theory, a defining feature of ideology is that it involves preferring the first-order preference ranking one actually holds to the alternatives. More specifically, in my account strong ideology is viewed as a special case of morality: in addition to requiring its adherent to prefer to hold it, as a point of personal duty, an ideology also requires that its adherent prefer to spread the ideology by converting others.

Because the theory of ideology I provide uses notions such as meta-payoffs that may be unfamiliar, I begin in section 3.2 with an informal extended example. The core concepts in the theory are introduced with a limited amount of notation and mathematical definition as a way to account intuitively for features in the preferences held by adherents to an ascetic religion and their secular counterparts. This lays the groundwork for a formal and more general theory of ideology in section 3.3.

3.2 Motivating example: an ascetic religion

In this section, I provide an extended motivating example that illustrates the notion of ideology I wish to focus on. The motivating example uses some concepts – such as meta-payoffs – that will be formally defined in section 3.3, but to begin with my aim is to present a relatively informal account to aid intuition.

Suppose there are n individuals, each of whom decides whether or not to abide by a code of behaviour in their everyday lives. This code of behaviour is prescribed by an ascetic religion, according to which an individual should eschew physical possessions and worldly comforts other than those needed for survival, on the basis that doing so will help him or her achieve spiritual enlightenment. In the "game of life" Γ – which represents players' everyday lives and interactions – each player has action set $Z = \{0, 1\}$, where the action 0 represents *refraining*, which means consuming only plain food in modest quantities, wearing simple, rough-hewn clothes and keeping very few and basic personal possessions. The action 1, in contrast, repre-

sents *indulging*, i.e. acquiring and consuming goods and services without practising self-denial beyond consumption-smoothing. There are two preference types of player: $\Theta = \{0, 1\}$, where type 0 is a *devout* player and type 1 is a *secular* player. Devout players ($\theta = 0$) prefer to *refrain* ($z = 0$), whatever other players choose, while secular players ($\theta = 1$) always prefer to *indulge* ($z = 1$). However, a player's payoffs can depend on others' actions. For devout players, the greater the number of other players playing *indulge*, the greater the temptation they face from the visible consumption of goods and services they see taking place around them. Although devout players still find it optimal to resist temptation, they bear a psychological cost in doing so. Suppose for simplicity that secular players, on the other hand, have payoffs unaffected by others' actions in the game of life Γ.

More formally, let $\mathbf{z} = (z_i, \mathbf{z}_{-i}) \in Z^n$ denote an action profile and let $\theta = (\theta_i, \theta_{-i}) \in \Theta^n$ denote a type profile. Assume that devout players have utility

$$u_{dev}(0, \mathbf{z}_{-i}) = 1 - \frac{1}{n} \sum_{j \neq i} z_j \tag{3.1}$$

$$u_{dev}(1, \mathbf{z}_{-i}) = 0 \tag{3.2}$$

Note that the payoff a devout player receives from refraining is always strictly positive, as the psychological cost represented by the second term on the right hand side of (3.1) can be at most $\frac{n-1}{n} < 1$ in absolute magnitude. As her payoff from indulging ($z_i = 1$) is always zero, regardless of other players' actions, she finds *refrain* ($z_i = 0$) to be a strictly dominant action. To begin with, we will suppose that secular players have the payoffs

$$u_{sec}(0, \mathbf{z}_{-i}) = 0 \tag{3.3}$$

$$u_{sec}(1, \mathbf{z}_{-i}) = 1 \tag{3.4}$$

for any $\mathbf{z}_{-i} \in Z^{n-1}$. Clearly, a secular player has a strictly dominant action to *indulge* ($z_i = 1$). As both types of player have a strictly dominant action in this example, there is a

unique Nash equilibrium, denoted $\mathbf{z}^*$, where [60]

$$z_i^* = \begin{cases} 0 & if\ i\ is\ devout \\ 1 & if\ i\ is\ secular \end{cases}$$

Now suppose that before they choose whether to refrain or indulge in the game of life Γ, players can attempt to influence each other's type. In particular, each player chooses whether to *invest* or *not invest*. Investment by a devout player involves preaching and otherwise attempting to convert the secular players they see around them to become followers of the religion. Investment by a secular player, similarly, might involve talking to devout players and questioning the basis for their beliefs. Suppose for simplicity that there is the following *conversion technology*:

- If player i invests, she will not change type – her investing is a way of consolidating her own stance on religious matters.
- If i does not invest, then she changes type only if more other-type players invest than same-type players.

Investing can thus convert others, in addition to immunising oneself from being converted. Finally, suppose that investing incurs a cost of $c \geq 0$ regardless of the player's type.

To analyse equilibrium play in this situation, we need to ask: would a devout player i be better off if her type changed? At first glance, the answer may appear to be Yes: secular players' payoffs are independent of other players' actions, and so they can guarantee a payoff of 1 by not investing and indulging ($z = 1$). In fact however, the question is as yet indeterminate. Consider the payoffs

[60] Throughout this chapter, I make the simplifying assumption (formalised in section 3.3) that Γ has a unique Nash equilibrium for any given type profile.

$$u^{'}(0, \mathbf{z}_{-i}) = -4 \tag{3.5}$$

and

$$u^{'}(1, \mathbf{z}_{-i}) = -3 \tag{3.6}$$

for any $\mathbf{z}_{-i} \in \{0,1\}^{n-1}$. As $u^{'}(.)$ is a positive affine transformation of $u_{sec}(.)$, it represents the same set of preferences. But now secular players can only receive negative payoffs. The problem is one of incomplete specification: to be able to assess whether a player would wish to change type, we need to know how players compare payoffs *across types*. Player i's comparison of payoffs across types is encoded by her *meta-payoffs*, which are von Neumann-Morgernstern (vNM) utilities as a function of outcome profiles Z^n and preference type profiles θ^n. To avoid confusion, I will refer to her payoffs defined simply over the set of outcome profiles Z^n, such as those in (3.1) to (3.6), as *first-order payoffs*.

An arbitrary meta-payoff for player i is denoted $w_i(z_i, \mathbf{z}_{-i}, \theta_i, \theta_{-i})$. For the case that i is secular, let us suppose her meta-payoffs are as follows.

$$w_{prag\,sec}(0, \mathbf{z}_{-i}, 0, \theta_{-i}) = 1 - \frac{1}{n}\sum_{j \neq i} z_j \tag{3.7}$$

$$w_{prag\,sec}(1, \mathbf{z}_{-i}, 0, \theta_{-i}) = 0 \tag{3.8}$$

$$w_{prag\,sec}(0, \mathbf{z}_{-i}, 1, \theta_{-i}) = 0 \tag{3.9}$$

$$w_{prag\,sec}(1, \mathbf{z}_{-i}, 1, \theta_{-i}) = 1 \tag{3.10}$$

for any $\mathbf{z}_{-i} \in Z^{n-1}$ and any $\theta_{-i} \in \Theta^{n-1}$, where $w_{prag\,sec}(0, \mathbf{z}_{-i}, 0, \theta_{-i}))$ denotes the (meta-) payoff that a secular player assigns to the outcome $(0, \mathbf{z}_{-i})$ *on the counterfactual that she were instead devout* (i.e. $\theta_i = 0$), and so on. Comparing equations (3.7) and (3.10), we see that a secular player is indifferent between switching to be a devout player (with $\theta_i = 0$) subject to the outcome she would then regard as the best (i.e. all players refraining), and continuing to be secular ($\theta_i = 1$) subject to the outcome she regards as the best (i.e. any profile in which she

indulges). Equally, from (3.8) and (3.9), a secular player is indifferent between switching to be devout subject to the outcome she would then regard as the worst (i.e. all players indulging), and continuing to be secular subject to the outcome she regards as the worst (i.e. any action profile in which she refrains). On this specification, a secular player is *pragmatic* in evaluating whether she would prefer to change types, placing equal value on her future self regardless of what preferences that self might possess.

While secular players adopt this pragmatic approach to comparing payoffs across types, devout players evaluate the possibility of changing type differently. To the devout ascetic, adherence to the religion is a core part of her identity, which she should seek to preserve regardless of whether she would find life easier, given others' behaviour, were she to change type.[61] In particular, she would prefer to remain devout even if everyone else around her were indulging and she were subject to temptation. One possible suitable specification of meta-payoffs for devout players is thus

$$w_{weak\,dev}(0, \mathbf{z}_{-i}, 0, \theta_{-i}) = 1 - \frac{1}{n}\sum_{j \neq i} z_j \tag{3.11}$$

$$w_{weak\,dev}(1, \mathbf{z}_{-i}, 0, \theta_{-i}) = 0 \tag{3.12}$$

$$w_{weak\,dev}(0, \mathbf{z}_{-i}, 1, \theta_{-i}) = -2 \tag{3.13}$$

$$w_{weak\,dev}(1, \mathbf{z}_{-i}, 1, \theta_{-i}) = -1 \tag{3.14}$$

for any $\mathbf{z}_{-i} \in Z^{n-1}$ and any $\theta_{-i} \in \Theta^{n-1}$. Under this specification, we see that a devout player would always rather remain devout, as the minimum meta-payoff she assigns to being devout (of zero) is greater than the maximum meta-payoff she assigns to being secular (of -1). We say that devout players with meta-payoffs of this sort are *ideological*. The ascetic religion can be thought of as an ideology.

The specification of devout players' meta-payoffs above assumes that they do not care

[61] If an attribute – in this case a preference type – forms part of one's identity, then seeking to preserve it can be thought of as akin to self-preservation.

directly about other players' preference types. Formally, the meta-payoffs $w_{weak\,dev}(.)$ are independent of θ_{-i}. In general, if i's payoffs are (a) ideological and (b) independent of θ_{-i}, we say they are *weakly ideological*. This may be reasonable if, for instance, preference types are not themselves observable.[62] Another reason the above specification may be reasonable is simply that the religion could be focused on the self; the individual is concerned with her own spiritual enlightenment, not that of others. Note that a player with the meta-payoffs $w_{weak\,dev}(.)$ as specified in equations (3.11) to (3.14) does nonetheless have an incentive to convert other players, since assuming all players play their dominant actions in Γ, she will receive a higher meta-payoff (3.11) if there are fewer secular players j playing $z_j = 1$. This incentive is known as an *instrumental incentive to convert*.

For comparison, consider the alternative specification under which devout players have the meta-payoffs

$$w_{str\,dev}(z_i, \mathbf{z}_{-i}, \theta_i, \theta_{-i}) = w_{weak\,dev}(z_i, \mathbf{z}_{-i}, \theta_i, \theta_{-i}) - \frac{1}{n}\sum_{j \neq i} \theta_j \tag{3.15}$$

In other words, aside from evaluating the outcome $\mathbf{z}$ of the game of life Γ and evaluating their own preference type θ_i, devout players also attach value directly to the preference types of other players in the profile θ_{-i}. Specifically, they attach a disutility of $\frac{1}{n}$ for every one of the other players who are secular. If devout players have meta-payoffs of this sort – or more generally have meta-payoffs under which they are (a) ideological and (b) for any fixed actions $(z_i, \mathbf{z}_{-i})$ and number of players n, their meta-payoffs increase in the number of same-type (i.e. devout) opponents – we say they are *strongly ideological*. Note that a strongly ideological devout player with payoffs given by (3.15) thus benefits in two ways if she converts her secular opponents to become devout. First, she benefits directly from a reduction in the magnitude of

[62] If preference types are not observable, however, the analysis is complicated somewhat without yielding much extra insight, so I assume complete information throughout. Players may be able to discern whether players are secular from conversations, say, or from observing that they do not follow any customs of the ascetic religion that are in place in addition to the code of behaviour that requires restraint in consumption.

the penalty term $\frac{1}{n}\sum_{j\neq i}\theta_j$. This effect is known as player i's *ideological incentive to convert*. Second, if play continues in equilibrium she benefits from an improvement in outcome in Γ, via a reduction in the magnitude of the term $\frac{1}{n}\sum_{j\neq i} z_j$ in (3.11). As before, the improvement in payoff she receives in Γ is known as her *instrumental incentive to convert*. In having both an ideological and instrumental incentive to convert, strongly ideological devout players differ from weakly ideological devout players, whose only incentive to convert is instrumental.

A final point to note is that we can relate the concept of ideology to the concept of homophily, i.e. the preference for sharing attributes in common with one's neighbours. If someone prefers that others share her preference type, even if such a change would move her from the best to the worst outcome of Γ, she displays homophily, a notion I define formally in section 3.3. By inspection of the payoff functions for each of the preference types above, only the strongly ideological players display homophily, because only they care directly about the preference types of the other players.

Now that meta-payoffs are specified, it is possible to analyse equilibrium play. Suppose that before investments take place, n_d players are devout and $n_s = n - n_d$ are secular. Recall that in the round of simultaneous investments before Γ is played, if i *invests* then she incurs a personal cost of $c \geq 0$. Investing shores up her own ideological position, ensuring that she will not change type. If she does not invest, however, then she changes type only if more other-type players invest than same-type players.

Case 1

First, consider the case where:

- devout players possess the weakly ideological meta-payoffs $w_{weak\,dev}(.)$; and
- secular players possess the pragmatic meta-payoffs $w_{prag\,sec}(.)$.

In this case, consider the decision facing a secular player. Recall that secular players are pragmatic, so they are indifferent between being either preference type while receiving that

type's best outcome in Γ (and similarly with respect to types' worst outcomes). A secular player i derives a meta-payoff of $1 - \frac{1}{n}\sum_{j \neq i} z_j$ from being converted to being devout and playing *refrain* ($z_i = 0$), the dominant action for a devout player, compared to a meta-payoff of 1 from remaining secular and playing her dominant action of *indulge* ($z_i = 1$). Note, however, that if i is converted, then all other secular players who do not invest will also be converted. Consequently, if no secular players invest, they will all jointly find not investing strictly optimal, as they will receive their maximum meta-payoff of 1, compared with a net meta-payoff $1-c$ from investing. Turning to devout players, from (3.11) we see that the benefit to a devout player from the conversion of all the secular players is $\frac{n_s}{n}$, as converted secular players j will play $z_j = 0$ as their dominant action. Thus, provided the cost of investing $c < \frac{n_s}{n}$, no devout player has a profitable deviation if precisely one devout player invests. There is then a set of n_d subgame perfect equilibria in which precisely one devout player invests, no secular players do, all secular players are converted to being devout and players then play their dominant action of *refrain* in the game of life Γ.[63] If $c > \frac{n_s}{n}$, on the other hand, then the only equilibrium is one in which no player invests.

These are in fact the only pure Nash equilibria. To see why, first consider the case where $n_s < n_d$, and suppose *a contrario* that some strictly positive number $k \leq n_s$ of the secular players invest in equilibrium. A mutual best response for devout players is if precisely $k+1$ of them invest, provided $c < \frac{n_s}{n}$.[64] Otherwise, they best-respond if none invest. In either case, now taking devout players' investments as the fixed portion of the investment profile, secular players' best response is for all to choose not to invest. It follows that no secular player invests in equilibrium if $n_s < n_d$. Turning to the case where $n_s \geq n_d$, we also need to consider the

[63] In general I restrict attention to pure strategies. If we allow for mixed strategies in this instance, however, if $c < \frac{n_s}{n}$ there is also clearly a symmetric subgame perfect equilibrium in which each devout player invests with probability p, where p is the unique root of the equation $\frac{n_s}{n}(1-p)^{n_d-1} = c$ subject to the constraint $p \in [0,1]$. The left hand side of this equation is the product of the benefit from converting the secular players of $\frac{n_s}{n}$ and the probability that a devout player's investment is pivotal of $(1-p)^{n_d-1}$ If this product equals the cost c of investing then devout players are indifferent between investing and not, which is required for a mixed strategy to be weakly optimal.

[64] Another mutual best response is if none invest, i.e. devout players fail to coordinate, which is also the (unique) best reponse when $c \geq \frac{n_s}{n}$.

case where k secular players invest such that $n_d < k \leq n_s$. In this case, all devout players will be converted regardless of their investment choice, and so all best-respond by not investing. But then any secular player finds a unilateral deviation strictly optimal, and so it cannot be that k secular players invest in equilibrium.

Case 2

Second, consider the case where:

- devout players possess the strongly ideological meta-payoffs $w_{str\,dev}(.)$; and
- secular players possess the pragmatic meta-payoffs $w_{prag\,sec}(.)$.

Equilibrium analysis follows the same lines as in the previous case, except that now the benefit to a devout player from the conversion of all the secular players is $\frac{2n_s}{n}$, twice its previous value. The minimum cost needed to ensure that no player invests is thus $\frac{2n_s}{n}$.

Case 3

Finally, let us examine the possibility that secular players too could be ideological. Their ideological meta-payoffs could perhaps reflect a conviction that the philosophy of self-denial is inherently wrong, and that asceticism goes against the natural order. However, when playing the game of life Γ, their payoffs are as before; devout players suffer a cost in resisting temptation if other players indulge, whereas secular players are indifferent to others' actions. In particular, the meta-payoffs of secular players are now as follows.

$$w_{weak\,sec}(0, \mathbf{z}_{-i}, 0, \theta_{-i}) = -1 - \frac{1}{n}\sum_{j \neq i} z_j \tag{3.16}$$

$$w_{weak\,sec}(1, \mathbf{z}_{-i}, 0, \theta_{-i}) = -2 \tag{3.17}$$

$$w_{weak\,sec}(0, \mathbf{z}_{-i}, 1, \theta_{-i}) = 0 \tag{3.18}$$

$$w_{weak\,sec}(1, \mathbf{z}_{-i}, 1, \theta_{-i}) = 1 \tag{3.19}$$

Note that these meta-payoffs assign a maximum and minimum possible values of -1 and -2 respectively to the case where the secular player changes type to become devout (i.e. switches from $\theta_i = 1$ to $\theta_i = 0$). In this way, they mirror the calibration of payoffs across types of the devout player; there is an effective penalty of absolute size 2 associated with a change in type in either case. This ensures that the payoffs are ideological. They are also clearly weakly ideological, since they do not depend on any θ_j for $j \neq i$. In addition to supposing secular players have these weakly ideological meta-payoffs, I suppose that devout players possess the weakly ideological meta-payoffs $w_{weak\,dev}(.)$.

It turns out that for intermediate values of the cost in case 3, multiple investment levels are possible in equilibrium. (For a more detailed equilibrium analysis of case 3, see Appendix C.2.) This is driven by the fact that players of either type, being ideological, have an ideological incentive to retain their own type, while players of at least one type (in this case, devout players) have an instrumental incentive to convert their opponents. If the cost c of investment is not too high, there is thus a ratchet effect: devout players wish to invest until their combined investment just exceeds that of the secular players, whereas secular players – who have no instrumental incentive – wish to invest until their combined investment matches that of the devout players. If there are at least as many secular players as devout players, then once all devout players invest, the secular players match them and an equilibrium is reached. Clearly, the higher the level of investment, the more inefficient the equilibrium, as the costly investment in these equilibria does not affect either the outcome of Γ or the type profile. If, however, the number of devout players exceeds that of secular players, then the ratchet process collapses into a cycle: once the secular players are unable to match the number of devout players, their best response is not to invest at all.

How general are these results? Section 3.3 below investigates ideological games in greater generality. It turns out that regardless of the first-order payoffs in Γ, if there are two ideological types, any pure-strategy equilibrium will either be one in which exactly one player invests, or will be an equilibrium described in case 3. As ever, if costs are prohibitive for both types,

neither will invest. For any cost beneath this level, if the more numerous type finds it optimal to convert the other type, either the other type finds it optimal to preserve their type (so there is no pure-strategy equilibrium) or not (in which case, one player of the former type invests). If there are the same number of players of each type, or if the more numerous type finds it optimal to preserve their type but not to convert, then there can be inefficient equilibria in which investment by players on both sides "cancels out".

If one type is pragmatic and the other ideological, on the other hand, and if the ideological type has equilibrium payoffs in Γ that are maximised when all players are of the ideological type (in which case the type is known as a *perfectly ideological* type) the results will be as in case 1 or case 2. This time the result is driven by the conversion technology we have assumed, whereby all players of a type are converted if and only if strictly more of the other-type players invest. Crucially, if a perfectly ideological type finds it optimal to convert a pragmatic type and does this successfully, then pragmatic players will enjoy their maximum meta-payoff since all players will receive their maximum first-order payoff as the ideological type in Γ. If, in contrast, the ideological type has equilibrium payoffs in Γ that are not maximised when all players are of that type, it is possible that the pragmatic type will find it optimal to preserve their type, or may even have an (instrumental) incentive to convert the ideological type, if their equilibrium first-order payoffs in Γ are improved by having other players of their own type. At the same time, it is also possible that the ideological type may not wish to convert the pragmatic players, though by the definition of ideology they will want to preserve their own type.

In the next section, I provide a formal account of ideology, building on the intuition from the extended example above, by identifying a class of finite games known as "ideological games". In this framework, I then provide a decomposition of the incentives faced by players with ideological preferences and use this to characterise sets of equilibria, at once generalising and formalising the results of section 3.2.

3.3 Formal model

3.3.1 Ideological games

An ideological game has two stages. Let $N = \{1, \dots n\}$ be the set of players. Play is as follows.[65]

Initial (first) stage. Before play begins, each player has a preference type known as her *initial type*; player i's initial type is denoted $\theta_i^{(1)} \in \Theta$, where Θ is a finite set of (at least two) types, and the initial type profile is denoted $\theta^{(\mathbf{1})} \in \Theta^n$. In the first stage, the players simultaneously choose whether to invest. Player i's investment choice is denoted $x_i \in \{0, 1\}$; the profile of investment choices is denoted $\mathbf{x} = (x_i, \mathbf{x}_{-i})$.[66] Define

$$X_\theta(\mathbf{x}) := \sum_{\{j \in N : \theta_j^{(1)} = \theta\}} x_j \tag{3.20}$$

X_θ is known as the *total investment by (type)* θ.

Final (second) stage. Immediately following investment choices in the first stage, players' preference types can change; a player's type in stage 2 is called her *final type*. Player i's final type, $\theta_i^{(2)} \in \Theta$, is determined by the investments that i and j make, as follows.

Assumption 3.1 *(Conversion technology) The final type profile* $\theta^{(\mathbf{2})} = f(\theta^{(\mathbf{1})}, \mathbf{x})$*, where the function* $f : \Theta^n \times \{0, 1\}^n \to \Theta^n$ *is defined as follows.*

[65] Time is not included in the model and throughout this chapter players do no discounting. It is natural to think of investments as leading to *future* payoffs in the second stage, however, and so I use time-related language in interpreting the features of the model and the results. Allowing for discounting of final-stage payoffs in the first stage would not lead to any substantive changes to the model's results, as it would be equivalent merely to a shift in the cost of investment c.

[66] An alternative approach to modelling investment in this context would include a second dimension, so that a player chooses both whether to 'attack' - that is, to attempt to influence others – and to 'defend' – that is to make oneself less prone to being influenced. In my model, a player's choice of investment is along one dimension. In support of this simplification, I note that drawing others' attention to a line of argument, or exposing them to advertising or propaganda, say, may often involve focusing one's own attention on the same object.

$$f(\theta^{(\mathbf{1})}, \mathbf{x}) = \begin{cases} \theta^{(\mathbf{1})} & if\ \neg\exists\theta'\ such\ that\ \forall\theta \neq \theta',\ X_{\theta'} > X_{\theta} \\ (\theta', \ldots, \theta') & if\ \exists\theta'\ such\ that\ \forall\theta \neq \theta',\ X_{\theta'} > X_{\theta} \end{cases} \tag{3.21}$$

The function $f(.)$ is known as the *conversion technology*. Assumption 3.1 implies that if one type of player invests strictly more than any other type of player, then all players convert to that type; otherwise, all players retain their initial types. The fact that the conversion technology takes this deterministic form simplifies the analysis below.[67]

Once the final type profile $\theta^{(\mathbf{2})}$ has been determined, the players simultaneously play a finite, one-shot game Γ. Purely to economise on notation, assume Γ is symmetric and let Z denote the (finite) action set of Γ, with $z_i \in Z$ denoting an arbitrary action. The game is one of complete information. In particular, the initial type profile $\theta^{(\mathbf{1})}$ is assumed to be common knowledge among players, as are investment choices and the final type profile (once they are realised). The game Γ is assumed to satisfy the following simplifying assumption.[68]

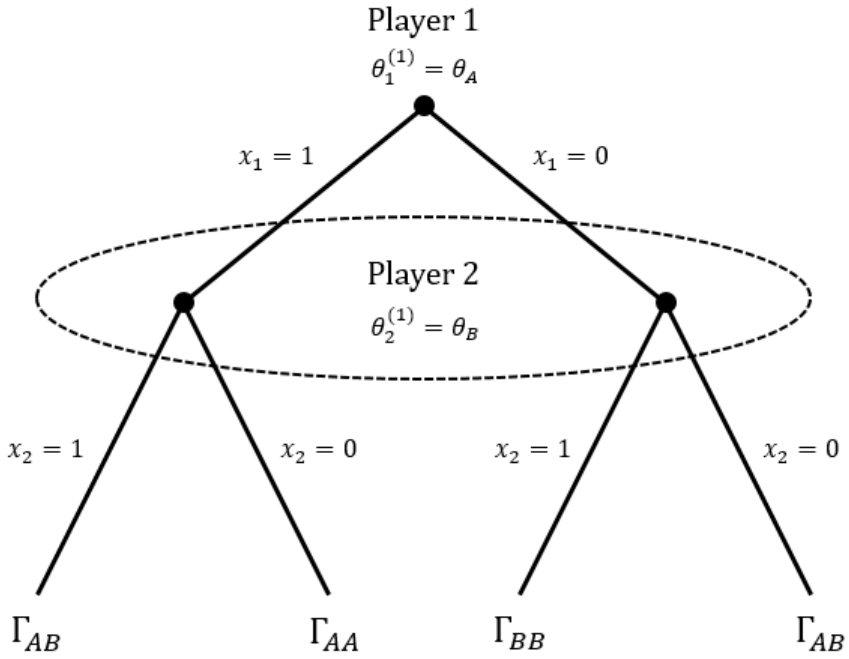
Assumption 3.2 *(Unique Nash equilibrium) Given any final type profile $\theta^{(\mathbf{2})} \in \Theta^n$, the continuation of Γ has a unique Nash equilibrium.*

Assumption 3.2 thus ensures that the final type profile $\theta^{(\mathbf{2})}$ determines a unique final-stage action Nash profile $\mathbf{z}^*(\theta^{(\mathbf{2})})$.

[67] However, the results of this section and elsewhere can also be obtained, for instance, via a conversion technology in which a player of initial type θ has an independent probability of switching to any given type $\theta' \neq \theta$ that is linearly increasing in $X_{\theta'}$. A sufficient condition for Proposition 3.2 and Theorem 3.1, for example, would then be that meta-payoffs be linear in the numbers of each type in the final type profile. Another point to note is that the conversion technology in the extended motivating example, according to which an investment by i immunises her from being converted, yields the same equilibria as Assumption 3.1. This is especially clear in the case of a two-player game, where the two technologies are exactly the same, as in equation (C.1) in Appendix C.3.

[68] A weaker but slightly less simple assumption than Assumption 3.2 that would have the same effect is that if a continuation of Γ contains more than one Nash equilibrium, one of them is specified as always being selected.

Figure 17: Sketch of extensive form for case of two players with different initial types



Example 3.1 *Suppose that* $\Theta = \{\theta_A, \theta_B\}$*, there are two players and the initial type profile is* $\theta^{(\mathbf{1})} = (\theta_A, \theta_B)$*. In this case, the subgame of* Γ *in which the players have final type profile* $\theta^{(\mathbf{2})} = (\theta_A, \theta_B)$ *is denoted* Γ_{AB}*, and so on. The extensive form of the game is then that sketched in Figure 17.*

□

Preferences. Let $\succeq$ be any (first-order) preference relation over ΔZ^n, the space of simple lotteries over outcomes of Γ.[69] A compound lottery over a pair of simple lotteries L_1 and L_2 is denoted simply as a convex combination of the two. Assume further that $\succeq$ meets the axioms of von Neumann-Morgernstern (vNM) rationality, as follows.

Assumption 3.3 *(First-order continuity) For any three lotteries* L_1*,* L_2*,* $L_3 \in \Delta Z^n$*, if* $L_3 \succeq L_2 \succeq L_1$*, then* $\exists \alpha \in [0,1]$ *such that* $\alpha L_1 + (1-\alpha)L_3 \sim L_2$

[69]By definition, a preference relation is a complete and transitive binary relation. I assume the relation does not entail indifference between all lotteries.

Assumption 3.4 *(First-order independence) For any three lotteries* L_1, L_2, $L_3 \in \Delta Z^n$, *if* $L_2 \succeq L_1$, *then* $\forall \alpha \in [0,1]$, $\alpha L_2 + (1-\alpha)L_3 \succeq \alpha L_1 + (1-\alpha)L_3$

By the vNM Utility Theorem, the preference relation $\succeq$ can thus be represented by vNM payoffs. With every $\theta \in \Theta$ there is associated a unique vNM first-order preference relation $\succeq_\theta$ over ΔZ^n. Let us denote by $u_\theta(.) : Z^n \to \mathbb{R}$ a utility function that represents $\succeq_\theta$.

Now let $\succeq^*$ be any preference relation (known as a *meta-preference relation*) over the domain $\mathcal{L}^* := \Delta[Z^n \times \Theta^n]$, the space of lotteries over the Cartesian product of the outcomes of Γ and the set of type profiles, with arbitrary element $\lambda \in \mathcal{L}^*$. Assume that $\succeq^*$ satisfies the following rationality axioms.

Assumption 3.5 *(Meta continuity) For any three lotteries* λ_1, λ_2, $\lambda_3 \in \mathcal{L}^*$, *if* $\lambda_3 \succeq^* \lambda_2 \succeq^* \lambda_1$, *then* $\exists \alpha \in [0,1]$ *such that* $\alpha\lambda_1 + (1-\alpha)\lambda_3 \sim^* \lambda_2$

Assumption 3.6 *(Meta independence) For any three lotteries* λ_1, λ_2, $\lambda_3 \in \mathcal{L}^*$, *if* $\lambda_2 \succeq^* \lambda_1$, *then* $\forall \alpha \in [0,1]$, $\alpha\lambda_2 + (1-\alpha)\lambda_3 \succeq^* \alpha\lambda_1 + (1-\alpha)\lambda_3$

With every $\theta \in \Theta$ there is associated a unique meta-preference relation $\succeq^*_\theta$ over $\mathcal{L}^*$, which can be represented by vNM payoffs. Formally, there exists a function $w : Z^n \times \Theta^n \to \mathbb{R}$ that maps from *2n*-tuples $(z_i, \mathbf{z}_{-i}; \theta_i^{(2)}, \theta_{-i}^{(2)})$ to the reals, unique up to a positive affine transformation, such that for any two lotteries λ_1, $\lambda_2 \in \mathcal{L}^*$, $\lambda_1 \succeq^* \lambda_2$ if and only if $E[w(\lambda_1)] \geq E[w(\lambda_2)]$.

The function $w : Z^n \times \Theta^n \mapsto \mathbb{R}$ is known as a *meta-utility function*. The numbers $w(z_i, \mathbf{z}_{-i}; \theta_i^{(2)}, \theta_{-i}^{(2)})$ are known as *meta-payoffs*, which are the second-order analogue of the vNM representation of (first-order) preferences over a standard set of outcomes or states. The completeness and transitivity properties of $\succeq^*$ together with Assumptions 3.5 and 3.6 are necessary and sufficient for the representation.

In summary, every type $\theta \in \Theta$ is thus characterised by a unique first-order preference relation $\succeq_\theta$, which can be represented by a vNM utility function $u_\theta(.)$, and a unique meta-preference relation $\succeq^*_\theta$, which can be represented by a vNM utility function $w_\theta(.)$.

To simplify the study of meta-preferences, I will restrict the class of admissible meta-utility functions by appealing to a notion of consistency between a decision-maker's first-order and meta-preferences. I wish to rule out certain combinations of first-order preference relations $\succeq_\theta$ and meta-preference relations $\succeq^*_\theta$. Consider, for example, a player's rankings over the outcomes of Γ supposing that she retains her current type when playing Γ. Her first-order preference relation $\succeq_\theta$ ranks lotteries over the outcomes of Γ, as does her meta-preference relation $\succeq^*_\theta$. If these rankings are inconsistent, then the type θ is, in a sense, ill-defined, for both rankings specify how she evaluates outcomes as an agent of type θ. For θ to be a meaningful type, it is therefore necessary to rule out inconsistency of this kind.[70]

To this end, for arbitrary lottery $\lambda \in \mathcal{L}^*$, denote by $\lambda^{actions} \in \Delta Z^n$ the marginal distribution of λ over Z^n, denote by $\lambda^{types} \in \triangle\Theta^n$ the marginal distribution of λ over Θ^n.

Assumption 3.7 *(Independence of preferences over action profiles and over others' types) For any player $i \in N$, let $\lambda_1, \lambda_2 \in \mathcal{L}^*$ be two lotteries in which player i has type θ with probability 1. If $\lambda_1^{types} = \lambda_2^{types}$, then $\lambda_1 \succeq^*_\theta \lambda_2$ iff $\lambda_1^{actions} \succeq_\theta \lambda_2^{actions}$.*

We then have the following Lemma.

Lemma 3.1 *(Identifying first-order with meta payoffs) For any player i of some type $\theta \in \Theta$, fix a representation $u_\theta(\mathbf{z})$ of $\succeq_\theta$. Then there exist (i) a representation $w_\theta(.)$ of $\succeq^*_\theta$ and (ii) a*

[70] An obvious stronger assumption on marginal distributions would be as follows. Let $\lambda_1, \lambda_2 \in \mathcal{L}^*$ be two lotteries where $\lambda_1^{types} = \lambda_2^{types}$ and $\lambda_1^{actions} = \lambda_2^{actions}$. Then for any type $\theta \in \Theta$, $\lambda_1 \sim^*_\theta \lambda_2$. In other words, a decision-maker is indifferent between any two lotteries that have the same marginal distribution over actions and the same marginal distribution over types. Such an assumption implies that for any player i of some type $\theta \in \Theta$, any representation $w_\theta(\mathbf{z}; \theta_i, \theta_{-i})$ of $\succeq^*_\theta$ can be expressed in the form $w_\theta(\mathbf{z}; \theta_i, , \theta_{-i}) = \tilde{v}_\theta(\mathbf{z}) + v_\theta(\theta_i, \theta_{-i})$, where the function $\tilde{v}_\theta(.) : Z^{n-1} \to \mathbb{R}$ maps from the set of outcomes of Γ to the reals and the function $v_\theta(.) : \Theta^{n-1} \to \mathbb{R}$ maps from the set of profiles of other players' types to the reals. The proof is analogous to that of Assumption 3.7. This alternative assumption is implausibly restrictive, however, as it implies that an agent derives utility from the outcomes of Γ depending only on her own initial type, and not on the final type profile. This would rule out, for instance, the three cases analysed in section 3.2, as all the meta-preference types considered have meta-payoffs that depend on other players' actions in Γ only if the player's final type is devout. The reason that all the meta-preferences types considered in section 3.2 have this feature arises from the specification of first-order preferences. In other words, devout players, by definition, care about other players' actions in Γ, whereas secular players do not. On the other hand, if all types in Θ are pragmatic, as defined by equations (3.26) to (3.29), then players' preferences meet the stronger assumption on marginal distributions.

function $v_\theta(.) : \Theta^{n-1} \to \mathbb{R}$ *mapping from the set of profiles of other players' types to the reals such that for any* $\mathbf{z} \in Z^n$, $w_\theta(\mathbf{z}; \theta, \theta_{-i}) = u_\theta(\mathbf{z}) + v_\theta(\theta_{-i})$.

Proof: Let $w_\theta(.)$ be a representation of $\succeq^*_\theta$. I establish the result by showing that (i) for any $\mathbf{z}, \mathbf{z}' \in Z^n$, the difference $w_\theta(\mathbf{z}; \theta, \theta_{-i}) - w_\theta(\mathbf{z}'; \theta, \theta_{-i})$ is independent of $(\theta_i, \theta_{-i}) \in \Theta^n$; and (ii) for any $\theta_{-i}, \theta'_{-i} \in \Theta^{n-1}$, the difference $w_\theta(\mathbf{z}; \theta, \theta_{-i}) - w_\theta(\mathbf{z}; \theta, \theta'_{-i})$ is independent of $\mathbf{z} \in Z^n$. Fix arbitrary $\mathbf{z}, \mathbf{z}' \in Z^n$, and $\theta_{-i}, \theta'_{-i} \in \Theta^{n-1}$. Let $\lambda_1 \in \mathcal{L}^*$ be the lottery in which $(\mathbf{z}; \theta, \theta_{-i})$ occurs with probability $\frac{1}{2}$ and $(\mathbf{z}'; \theta, \theta'_{-i})$ occurs with probability $\frac{1}{2}$. Let $\lambda_2 \in \mathcal{L}^*$ be the lottery in which $(\mathbf{z}'; \theta, \theta_{-i})$ occurs with probability $\frac{1}{2}$ and $(\mathbf{z}; \theta, \theta'_{-i})$ occurs with probability $\frac{1}{2}$. We have that $\lambda_1^{types} = \lambda_2^{types}$ and $\lambda_1^{actions} = \lambda_2^{actions}$, with the latter equality implying both $\lambda_1^{actions} \succeq_\theta \lambda_2^{actions}$ and $\lambda_2^{actions} \succeq_\theta \lambda_1^{actions}$. As in both lotteries, i is of type θ for sure, Assumption 3.7 then implies that $\lambda_1 \sim^*_\theta \lambda_2$. As $w_\theta(.)$ is a representation of $\succeq^*_\theta$, we have

$$\frac{1}{2} w_\theta(\mathbf{z}; \theta, \theta_{-i}) + \frac{1}{2} w_\theta(\mathbf{z}'; \theta, \theta'_{-i}) = \frac{1}{2} w_\theta(\mathbf{z}'; \theta, \theta_{-i}) + \frac{1}{2} w_\theta(\mathbf{z}; \theta, \theta'_{-i}) \tag{3.22}$$

i.e. (i) $w_\theta(\mathbf{z}; \theta, \theta_{-i}) - w_\theta(\mathbf{z}'; \theta, \theta_{-i}) = w_\theta(\mathbf{z}; \theta, \theta'_{-i}) - w_\theta(\mathbf{z}'; \theta, \theta'_{-i})$. Rearranging yields (ii) $w_\theta(\mathbf{z}; \theta, \theta_{-i}) - w_\theta(\mathbf{z}; \theta, \theta'_{-i}) = w_\theta(\mathbf{z}'; \theta, \theta_{-i}) - w_\theta(\mathbf{z}'; \theta, \theta'_{-i})$. □

The meta-utility function, unique up to a positive affine transformation, which represents $\succeq^*_\theta$ is denoted $w_\theta(.)$. Assumption 3.7 allows us to identify a type with a set of first-order payoffs *and* with a set of meta-payoffs in a consistent, meaningful way. It means that a player, when choosing whether to invest ex-ante, anticipates the first-order payoffs they will have in the game of life Γ and, conditional on these payoffs (i.e. aside from any preference for a particular final type profile) acts as an expected utility maximiser accordingly.

Initial stage payoffs. A player's *stage 1 (ex-ante) payoffs* represent the preferences a player possesses in stage 1 of the ideological game. Players suffer a cost $c \geq 0$ from investing (i.e. from playing $x = 1$ in the initial stage.) The *initial (ex-ante) payoffs* of a player i with a given initial type $\theta \in \Theta$ are as follows.

$$\pi_\theta \left(x_i; z_i, z_{-i}; \theta_i^{(2)}, \theta_{-i}^{(2)} \right) = -cx_i + w_\theta \left(z_i, z_{-i}; \theta_i^{(2)}, \theta_{-i}^{(2)} \right) \tag{3.23}$$

The first term on the right hand side of (3.23) reflects that all players, regardless of type, suffer a disutility of c in stage 1 if and only if they invest; additionally, they receive a second-order payoff, which is the second term on the right hand side. In other words, for a player of a given initial type, her initial stage payoffs depend on her own first-stage investment costs, the outcome of Γ in the final stage, her own initial type and the final type profile.[71]

Final stage payoffs. Recall that the utility function $u(.)$ specifies the payoffs over the outcomes of Γ for each final preference type.

Strategies and summary. Define the *set of histories* $H = \{0,1\}^n \times \Theta^{2n}$; an arbitrary element is denoted $h \in H$ and is called a *history.* A history is a tuple comprising an investment profile, initial type profile and final type profile; we can write it explicitly as $h = (\mathbf{x}; \theta^{(\mathbf{1})}; \theta^{(\mathbf{2})})$. A strategy is a pair $(x(.), z(.))$, where the *initial stage strategy* $x : \Theta \to \{0,1\}$ is a mapping from initial types to investment actions, and the *final stage strategy* $z : H \to \triangle Z$ is a mapping from histories to final stage actions.

The features of an ideological game set out above can be summarised in the following formal definition.

An *ideological game* is a tuple $\left\langle N, X, \Theta, \theta^{(\mathbf{1})}, c, f(.), Z, \{\succeq_\theta\}_{\theta\in\Theta}, \{\succeq^*_\theta\}_{\theta\in\Theta}\right\rangle$, where

- $N = \{1, \ldots, n\}$ is the set of players
- $X = \{0,1\}$ is the initial stage action set
- Θ is the finite set of types
- $\theta^{(\mathbf{1})} \in \Theta^n$ is the initial type profile
- $c \geq 0$ is the investment cost

[71] To see why it is necessary to define stage 1 payoffs in addition to stage 2 payoffs, note that the ex-post payoffs in stage 2 for either preference type – such as the set of payoffs $\{u_A(z)\}_{z\in Z}$ – are only unique up to a positive affine transformation. In order to model players as expected utility maximisers in stage 1 as well as in stage 2, it is thus necessary to 'fix' the sets of stage 2 payoffs for each final type relative to each other. In other words, some assumption about comparability of preferences across types is needed. This job is done by the second-order preference functions $w_A(.)$ and $w_B(.)$

- $f : \Theta^n \times \{0,1\}^n \to \Theta^n$ is the conversion technology
- Z is the (finite) final stage action set
- $\{\succeq_\theta\}_{\theta\in\Theta}$ are the first-order preference relations associated with each type
- $\{\succeq^*_\theta\}_{\theta\in\Theta}$ are the meta-preference relations associated with each type

The final-stage game Γ is given by the tuple $\langle N, Z, \Theta, \{\succeq_\theta\}_{\theta\in\Theta}\rangle$.

3.3.2 A taxonomy of meta-payoffs: homophily and ideology

In this section, I construct a categorisation of meta-payoffs in terms of two concepts: homophily and ideology, each with gradations of intensity. The formulation I give here of these concepts is not intended to be exhaustive or fine-grained; it is largely guided by the aim of a set of categories that readily permit equilibrium analysis. I begin with a short qualitative discussion of the concepts before moving on to formal definitions.

Homophily involves the tendency of individuals to seek to associate with others who display similar traits or preferences. It has been widely theorised and observed in a variety of settings, including learning and in the formation of friendships (Currarini, Jackson and Pin, 2009) and among those who share ideological material online on social networks (Bakshy et al, 2015). In the present context, the object of homophily will be (first-order preference) types. Below I consider the case of an agent who, conditional on maintaining her own type, prefers that her opponent share this type; this notion will then form the basis for a formal definition of Partial Homophily (PH). A natural way to strengthen the definition is then to suppose that whatever her final type, she prefers a same-type pairing; I formally define Full Homophily (FH) along these lines. Intuitively, an agent adhering to FH prefers to fit in, ahead of any concerns that her present type should proliferate; in this sense, there appears to be a tension between homophily and the notion of an ideologically-motivated agent.

The concept of ideology, as I will develop it here, focuses on two concerns. At its most basic level, an ideology requires its adherents to be motivated to retain their own preferences. In this respect, I work on the basis that all ideology has, by its nature, a defensive character.

This defensive character neither conflicts with nor requires homophily assumptions, since it is purely inward-looking. The second (and perhaps more obvious) concern of an ideology, which can be included as a condition to strengthen the concept, is that its adherents be motivated to persuade or convert others to one particular way of evaluating the world. A "strong" ideology of this sort is at odds with the notion of FH, since it requires an agent to prefer her opponent to convert to her initial type, even if she fears that she herself may be converted out of it. It does, however, imply PH, as it motivates agents to convert others to their own initial type.

Let us turn to a formal account of homophily. Define the *count* $\mathcal{C}_\theta(\Theta)$ of a type profile Θ with respect to a type $\theta \in \Theta$ to be the number of times θ appears in Θ. In addition to imposing Assumptions 3.3 to 3.7, let us assume the following.

Assumption 3.8 *(Impersonality) For any* $\mathbf{z} \in Z^n$*, any* $\theta \in \Theta$ *and any two* $\theta_{-i}, \theta'_{-i} \in \Theta^{n-1}$ *that have the same count with respect to every type,* $w_\theta(\mathbf{z}; ., \theta_{-i}) = w_\theta(\mathbf{z}; ., \theta'_{-i})$.

Assumption 3.8 says that while players may care directly about how many players are of each type, beyond that they do not care which players are of which type.

I now examine two alternative restrictions. The first alternative restriction, which is strongly intuitive and which is supported by empirical evidence such as that referred to above, is as follows.

Assumption 3.9 *(Partial Homophily) For any* $\mathbf{z} \in Z^n$ *and any* $\theta \in \Theta$*,* $w_\theta(\mathbf{z}; \theta, \theta_{-i})$ *is increasing in* $\mathcal{C}_\theta(\theta_{-i})$ *.*

The assumption of PH is simply that, on condition that a player maintains her type θ from the initial to the final stage, she prefers that her opponent share this type. The natural alternative to this assumption is to strengthen it to yield the following.

Assumption 3.10 *(Full Homophily) For any* $\mathbf{z} \in Z^n$ *and any* $\theta, \theta' \in \Theta$*,* $w_\theta(\mathbf{z}; \theta', \theta_{-i})$ *is increasing in* $\mathcal{C}_{\theta'}(\theta_{-i})$ *.*

Note that in this case, even in situations where a player of type θ changes to some other type $\theta^{'}$, her ex-ante preference (as type θ) is for her opponents to share her final type $\theta^{'}$. FH clearly implies PH, since it includes the case that $\theta^{'} = \theta$, which reduces to the definition of PH. Put simply, FH means that whatever a player's final type, she would rather that her opponent share that final type.

I now turn to categorising meta-payoffs via ideology. I first formally define three categories of meta-payoff: "ideological", "strongly ideological" and "pragmatic". I then discuss how to interpret the categories and consider the properties of each kind of ideological payoff under each alternative homophily assumption.

A player's meta-payoffs (and her accompanying meta-preferences) are *ideological*, with respect to her initial type θ, iff, whatever the other players' final type profile $\theta_{-i} \in \Theta^{n-1}$,

$$\min_{\mathbf{z} \in Z^n} w_\theta \left(\mathbf{z}; \theta, \theta_{-i}\right) > \max_{\mathbf{z} \in Z^n} w_\theta \left(\mathbf{z}; \theta^{'}, \theta_{-i}\right) \tag{3.24}$$

for any type $\theta^{'} \neq \theta$. In this case θ may be referred to as an *ideological type*, and players with ideological payoffs are themselves said to be ideological.

A player of initial type θ and the associated meta-payoffs $w_\theta(.)$ are *strongly ideological* if and only if (i) they are ideological and (ii) the following inequality of "Strong Ideology" (SI) holds, for any final type $\theta^{'} \in \Theta$ and for any type profiles for the other players $\tilde{\theta}_{-i}, \hat{\theta}_{-i} \in \Theta^{n-1}$ such that $\mathcal{C}_\theta(\tilde{\theta}_{-i}) > \mathcal{C}_\theta(\hat{\theta}_{-i})$.

$$\min_{\mathbf{z} \in Z^n} w_\theta \left(\mathbf{z}; \theta^{'}, \tilde{\theta}_{-i}\right) > \max_{\mathbf{z} \in Z^n} w_\theta \left(\mathbf{z}; \theta^{'}, \hat{\theta}_{-i}\right) \tag{3.25}$$

In other words, if SI holds, then ex-ante a player prefers that other players switch to her initial type even if she herself is converted to a different type (since $\theta^{'}$ may be different from the player's initial type θ). Note that SI implies PH; this is clear by considering the case that $\theta^{'} = \theta$, as inequality (3.25) then implies that for any given $\mathbf{z} \in Z^n$, $w_\theta \left(\mathbf{z}; \theta, \tilde{\theta}_{-i}\right) > w_\theta \left(\mathbf{z}; \theta, \hat{\theta}_{-i}\right)$ provided $\mathcal{C}_\theta(\tilde{\theta}_{-i}) > \mathcal{C}_\theta(\hat{\theta}_{-i})$, which is simply a restatement of PH. If a player is ideological but *not* strongly ideological, she is *weakly ideological*.

If, on the other hand, a player (with initial type θ) has meta-payoffs such that for any $\theta' \in \Theta$, any $\theta_{-i} \in \Theta^{n-1}$, any $\theta'_{-i} \in \Theta^{n-1}$ and any $\mathbf{z} \in Z^n$, the following four conditions hold:

$$w_\theta(\mathbf{z}; \theta, \theta_{-i}) = w_\theta\left(\mathbf{z}; \theta, \theta'_{-i}\right) \tag{3.26}$$

and

$$\max_{\mathbf{z} \in Z^n} w_\theta(\mathbf{z}; \theta, \theta_{-i}) = \max_{\mathbf{z} \in Z^n} w_\theta\left(\mathbf{z}; \theta', \theta_{-i}\right) \tag{3.27}$$

and

$$\min_{\mathbf{z} \in Z^n} w_\theta(\mathbf{z}; \theta, \theta_{-i}) = \min_{\mathbf{z} \in Z^n} w_\theta\left(\mathbf{z}; \theta', \theta_{-i}\right) \tag{3.28}$$

and

$$w_\theta\left(\mathbf{z}; \theta', \theta_{-i}\right) = w_{\theta'}\left(\mathbf{z}; \theta', \theta_{-i}\right) \tag{3.29}$$

then we say the player is *pragmatic*.

In the case of ideological payoffs as defined by (3.24), a player has an ideological type if in the initial stage she judges that no outcome of the final-stage game Γ obtained while being the other (rival) type is better than any outcome obtained while she possesses her initial type. Intuitively, agents are ideological if they would rather possess their current preferences regardless of what outcomes ensue when interacting with other people. For instance, a die-hard monarchist living in a proud republic might prefer the current world, in which he maintains his current beliefs but is ostracised by wider society, to one in which he himself becomes a republican and is embraced by one and all.[72] Although the agent can envisage changing his mind and outlook – and may even envisage being content in doing so – he does not currently value that future self as much as the one who shares his current preferences. Put simply, really believing in a principle requires one to hold true to it.

Additionally, an agent's meta-payoffs (which could reflect, for instance, the adoption of a moral or ideological principle) may also extend to the future selves of *other* agents. In the case of strongly ideological payoffs – inequality (3.25) – this entails preferring that one's

[72]This sort of situation would be represented by a game Γ in which a same-final-type pair of players enjoys a good Nash outcome (as ranked by that preference type).

opponent ends up adopting one's current preferences, regardless of what happens to oneself. Note that as defined here, although being strongly ideological implies preferring that one's future self have one's initial preference type (as having strongly ideological payoffs implies that the payoffs are ideological), it is consistent with the agent attaching more weight to her opponent's final type than his own. For instance, an ideological agent might prefer that her opponent adopt the agent's current type, even if this comes at the expense of her own future self having a different type. A reason for such a preference could be, say, that the opponent occupies a position of great importance or influence in society.[73]

In the case of pragmatic payoffs, condition (3.26) says that a pragmatic player does not care about other players' types directly, while conditions (3.27) and (3.28) say that in her ex-ante decision-making, such a player calibrates her sets of final-stage payoffs for each final type such that their ranges coincide. Condition (3.29) says that a pragmatic player of type θ, conditional on having final type θ', ranks outcomes of Γ in the same way as θ. Recall that Assumption 3.7 requires that any type θ, conditional on having final type θ', ranks outcomes of Γ via their meta-payoffs the same way as in their first-order payoffs. Condition (3.29) for a pragmatic type is effectively a generalisation of this assumption, requiring that pragmatic players derive ex-ante utility from a given outcome of Γ conditional on being converted to any type as they would do in the second stage having been so converted. Put simply, the idea is that a pragmatic player puts herself in the shoes of any other given type of player. Note that the reason conditions (3.27) and (3.28) are needed in addition to condition (3.29) is that without them, it could be the case that a pragmatic player ranked all outcomes of Γ conditional on being one type higher than all outcomes of Γ conditional on being another type.[74] Such a possibility goes against the notion of pragmatism, according to which a decision-maker should

[73] An ideologically-motivated spy, for instance, might find it worthwhile to try to "turn" senior figures in the enemy's government, even if he knew that, in so doing, he would be at risk of being "turned" himself, and becoming a double agent.

[74] Conditions (3.27), (3.28) and (3.29) therefore together require that for a pragmatic type to be well-defined, all types must have the same maximum and minimum meta-payoffs conditional on retaining their initial type. That this requirement can be achieved is clear from the fact that meta-payoffs are a unique representation of their respective meta-preferences up to a positive affine transformation, and so in specifiying an ideological game, we can set any two meta-payoffs for a given type to any required distinct values.

value all her future selves equally.

The definition of a pragmatic type, consisting in conditions (3.26) to (3.29), implies the following Lemma.

Lemma 3.2 *Suppose all types in* Θ *are pragmatic and let* i *be a player of an arbitrary pragmatic type* $\theta \in \Theta$. *Then there is a representation* $w_\theta(.)$ *of* $\succeq^*_\theta$ *and a representation* $u_{\theta'}(.)$ *of* $\succeq_{\theta'}$ *such that for any* $\theta' \in \Theta$, $w_\theta\left(\mathbf{z}; \theta', \theta_{-i}\right) = u_{\theta'}(\mathbf{z})$. *There exist numbers* $\underline{w}$ *and* $\overline{w} > \underline{w}$ *such that for any* $\theta' \in \Theta$, $\min\limits_{\mathbf{z} \in Z^n} u_{\theta'}(\mathbf{z}) = \underline{w}$ *and* $\max\limits_{\mathbf{z} \in Z^n} u_{\theta'}(\mathbf{z}) = \overline{w}$.

Proof: Let $\theta \in \Theta$ be a pragmatic type and suppose all types in Θ are pragmatic. Inequality (3.27) implies that a player of this type is indifferent between being any type whose best-ranked pair $(\mathbf{z}, \theta_{-i}) \in Z^n \times \Theta^{n-1}$ obtains. Call this payoff $\overline{w}$. Inequality (3.28) implies that a pragmatic player is indifferent between being any type whose worst-ranked outcome of Γ obtains. Call this payoff $\underline{w}$. Then for all $\theta' \in \Theta$, $\theta_{-i} \in \Theta^{n-1}$, and $\mathbf{z} \in Z^n$, a player of type θ has meta payoffs $w_\theta\left(\mathbf{z}; \theta', \theta_{-i}\right) \in [\underline{w}, \overline{w}]$. By Lemma 3.1, for every type $\theta' \in \Theta$, there is a first-order preference representation $u_{\theta'}(\mathbf{z}) = w_{\theta'}\left(\mathbf{z}; \theta', \theta_{-i}\right) - v_{\theta'}(\theta_{-i}) \in [\underline{w}, \overline{w}]$. By inequality (3.26), a pragmatic player is indifferent between different profiles of other players' types and so $v_{\theta'}(\theta_{-i})$ is a constant; set $v_{\theta'}(\theta_{-i}) = 0$, so $w_\theta(\mathbf{z}; \theta, \theta_{-i}) = u_\theta(\mathbf{z})$. By inequality (3.29), $w_\theta\left(\mathbf{z}; \theta', \theta_{-i}\right) = w_{\theta'}\left(\mathbf{z}; \theta', \theta_{-i}\right) = u_{\theta'}(\mathbf{z})$. □

Given any particular final stage outcome $\mathbf{z}$, among all possible pragmatic types, the pragmatic player of type θ most prefers to be that pragmatic type θ' whose first-order payoff $u_{\theta'}(\mathbf{z})$ takes the greatest value. Informally, we can think of pragmatic meta-payoffs as capturing a situation in which an agent has a particular set of (first-order) preferences, but has no attachment to those preferences in principle. In particular, it follows from inequalities (3.26) to (3.28) when a player evaluates among her possible future selves at the initial stage, she would simply rather become the future self for which z yields the greatest utility gain over that self's worst outcome.

The two alternative homophily assumptions imply different properties for ideologies, as follows.

Proposition 3.1 *(Homophily and ideology) Full homophily is inconsistent with strong ideology. Partial homophily is implied by strong and weak ideology.*

Proof: See Appendix C.1. To see why strong ideology and FH are inconsistent, consider a scenario in which a player of initial type θ prefers a type profile in which he is the only player of type θ to the profile in which he and all other players are of some other type θ'. This meta-preference, which is a necessary condition for a strong ideology, clearly implies the falsity of FH. Because the notion of strong ideology is theoretically appealing, allowing us to account for a desire to convert others, and because PH, not FH, is the strain of homophily for which there is empirical evidence, I henceforth assume PH.

The connection between homophily and ideology can be brought out most clearly in the case of two players of different types. If investment is cheap enough (i.e. $c \geq 0$ is sufficiently small), a strongly ideological player will always want to invest regardless of her opponent's investment level, since she both wants to maintain her own type and to convert her opponent. One key point of note is that even when investment is cheap, there may exist a no-invest equilibrium among a pairing of two different weakly ideological types, in addition to the both-invest equilibrium, provided that PH does not hold. The reason for this is that weakly ideological players, unlike strongly ideological ones, may be primarily defensive in their motivation. They seek to sustain their initial type profile without concerning themselves with converting others. There thus exists the possibility that two weakly ideological players, even though of different types, opt for peaceful coexistence.[75] Note, however, that the existence of multiple equilibria requires that both types be only "minimally" ideological: their only overriding ideological motivation is to maintain their own type. If, on the other hand, PH holds for at least one type and investment is cheap enough, both types invest in equilibrium, ruling out peaceful coexistence. Similarly, even if neither type satisfies PH, if we consider decreasing the cost

[75]The phase "peaceful coexistence" was famously used by Nikita Khrushchev during the Cold War (e.g. Khrushchev, 1959). Communist ideology in the Soviet Union had by the 1950s become primarily defensive in its outlook, in contrast to the earlier doctrine, promoted by Leon Trotsky, of "permanent revolution", by which the Soviet Union should attempt to foment revolutions in other other societies (Lerner, 1964). Peaceful coexistence between two ideologies, on my characterisation, can result only if both ideologies are weak; permanent revolution results from a strong ideology.

$c > 0$ of investment, then instrumental incentives – i.e. the possibility of investing to achieve a better outcome in Γ – may induce investment.

3.3.3 Equilibrium analysis

Under the assumption of complete information, the natural solution concept for an ideological game is that of subgame-perfect equilibrium. Recall that a strategy is a pair $(x(.), z(.))$, where $x : \Theta^n \to \{0, 1\}$ is a mapping from initial types to investment actions, and $z : H \to Z$ is a mapping from histories to final stage actions. A strategy $(x(.), z(.))$ is an *equilibrium* if and only if the following two conditions are met for all players $i \in N$, for any initial type profile $\theta^{(\mathbf{1})} \in \Theta$ and any history $h \in H = \{0, 1\}^n \times \Theta^{2n}$.

1. For every $z' \in Z$,

$$u_i\left(z_i(h), \mathbf{z}_{-i}(h)\right) \geq u_i\left(z', \mathbf{z}_{-i}(h)\right)$$

   where i's type is her final type in the history h.

2. For $x' \neq x_i$,

$$\begin{aligned} & w_i\left((x_i, x_{-i})(\theta^{(\mathbf{1})}), \mathbf{z}(h(x_i, x_{-i})\right) \\ \geq \; & w_i\left((x', x_{-i})(\theta^{(\mathbf{1})}), \mathbf{z}(h(x_i, x_{-i}))\right) \end{aligned}$$

where $h(x_i, x_{-i})$ is the history induced by investment choices (x_i, x_{-i}) according to the conversion technology $f(.)$ given the initial type profile $\theta^{(\mathbf{1})}$.[76] Throughout this section I take the initial type profile $\theta^{(\mathbf{1})}$ as arbitrarily fixed. The existence of an equilibrium is guaranteed by the fact that an ideological game, as defined here, has a finite set of players and a finite strategy set.

[76] I use the notational shortcut that $u_i(.) = u_A(.)$ and $w_i(.) = w_A(.)$ if $\theta_i^{(1)} = \theta_A$; $u_i(.) = u_B(.)$ and $w_i(.) = w_B(.)$ if $\theta_i^{(1)} = \theta_B$.

Recall that Assumption 3.2 ensures that any final type profile induces a unique Nash profile in the second stage. As equilibrium condition 1 requires that second stage actions form a Nash equilibrium, in characterising equilibria we can simply restrict attention to Markov final-stage strategy profiles $(z_i(\theta^{(\mathbf{2})}), z_{-i}(\theta^{(\mathbf{2})}))$, and in particular to the unique equilibrium final-stage strategy profile denoted $\mathbf{z}^*(\theta)$.

Denote by $\mathbf{x}_{-\theta}$ the (partial) investment profile among all players not of initial type θ. Define

$$\overline{X}_{-\theta} := \max_{\theta' \in \Theta} [X_{\theta'}(\mathbf{x}_{-\theta})] \tag{3.30}$$

recalling that $X_{\theta'}(\mathbf{x}_{-\theta})$ is the total investment by type θ' in $\mathbf{x}_{-\theta}$, known as the *maximum other-type investment* (relative to type θ). Finally, for an arbitrary fixed investment profile $\mathbf{x}$ define

$$\overline{X} := \max_{\theta' \in \Theta} [X_{\theta'}(\mathbf{x})] \tag{3.31}$$

The following Lemma will help simplify equilibrium analysis.

Lemma 3.3 *(Mutual best responses) Fix an arbitrary initial type profile $\theta^{(\mathbf{1})}$ containing at least one player of initial type θ, and fix a partial investment profile $x_{-\theta}$. Then in any mutual best response among type-θ players in pure strategies, either no type-θ players invest, or $\overline{X}_{-\theta}$ players do, or $\overline{X}_{-\theta} + 1$ players do.*

***Proof**:* No total investment level between zero and $\overline{X}_{-\theta}$ by type-θ players alters the outcome, so all such investing players would be strictly better off not investing. Likewise, any investment level above $\overline{X}_{-\theta} + 1$ will yield the same final type profile as if $\overline{X}_{-\theta}$ type-θ players invested, and so any such investing player would be strictly better off not investing. □

The final definitions needed for the purpose of characterising equilibria relate to the marginal benefit to investing. It will be notationally convenient to denote by $A \in \Theta$ an arbitrary type. First, define

$$\Delta_A^{convert} := w_A(\mathbf{z}^*(\mathbf{A}); \mathbf{A}) - w_A(\mathbf{z}^*(\theta^{(\mathbf{1})}); \theta^{(\mathbf{1})}) \tag{3.32}$$

where $\mathbf{A} \in \Theta^n$ is the final type profile in which all players are of type A. $\Delta_A^{convert}$ denotes the ex-ante net benefit to a type-A player from a final type profile (and the induced outcome $\mathbf{z}^*(\mathbf{A})$ of Γ) in which all her opponents are converted to her own type, compared with the status quo (i.e. the initial type profile and the outcome it would induce). Second, let $B \in \Theta$ be an arbitrary type and define

$$\Delta_{A,B}^{retain} := w_A(\mathbf{z}^*(\theta^{(\mathbf{1})}); \theta^{(\mathbf{1})}) - w_A(\mathbf{z}^*(\mathbf{B}); \mathbf{B})) \tag{3.33}$$

where $\mathbf{B} \in \Theta^n$ is the final type profile in which all players are of type B. $\Delta_{A,B}^{retain}$ denotes the ex-ante net benefit to a type-A player to retaining the status quo compared with a final type profile and outcome in which she and all other players are of type B. Note that if A is an ideological type, then by (3.24), $\Delta_{A,B}^{retain} > 0$.

We can relate $\Delta_A^{convert}$ and $\Delta_{A,B}^{retain}$ to players' investment incentives as follows. First, take an arbitrary player i of type A and fix a partial investment profile $\mathbf{x}_{-i}$ such that $X_A(\mathbf{x}_{-i}) = \overline{X}_{-A}$. Recall that $X_A(\mathbf{x}_{-i})$ is the total investment by type A in $\mathbf{x}_{-i}$, i.e. the total investment by players of type A excluding player i.[77] $\overline{X}_{-A}$ is the total investment by the type other than A that does the most investing. This means that type-A players other than i invest, in aggregate, as much as players of the most-investing type other than A. Given the investments $\mathbf{x}_{-i}$ of all players other than i, player i's investment decision is therefore pivotal between the status quo final type profile and and the all-A final type profile. $\Delta_A^{convert}$ can be expressed as follows.[78]

$$\Delta_A^{convert} := E\left[\pi_A(x_i = 1, \mathbf{x}_{-i})\right] - E\left[\pi_A(x_i = 0, \mathbf{x}_{-i})\right] + c \tag{3.34}$$

where $\mathbf{x}_{-i}$ is such that $X_A(\mathbf{x}_{-i}) = \overline{X}_{-A}(\mathbf{x}_{-i})$. The quantity $\Delta_A^{convert}$ is thus the individual

[77] (3.20) defines $X_A(\mathbf{x})$; substituting $\mathbf{x}_{-i}$ for $\mathbf{x}$ yields $X_\theta(\mathbf{x}_{-i}) := \sum_{\{j \in N \setminus \{i\} \,:\, \theta_j^{(1)} = \theta\}} x_j$

[78] (3.35) and (3.34) use shorthand notation, where e.g. $\pi_A(x_i = 1, \mathbf{x}_{-i})$ denotes $\pi_A(x_i, z_i, z_{-i}; \theta_i^{(2)}, \theta_{-i}^{(2)})$ given that $x_i = 1$, the final type profile $(\theta_i^{(2)}, \theta_{-i}^{(2)})$ is that implied by the initial type profile, the investment profile $(x_i, \mathbf{x}_{-i})$ and the conversion technology, and $(z_i, z_{-i}) = \mathbf{z}^*(\theta^{(\mathbf{2})})$.

marginal benefit from investing, neglecting the cost of investment c, to each type-A player when together type-A players invest marginally more (i.e. have one more unit of aggregate investment) than the next-most-investing type(s).

Second, fix a partial investment profile $\mathbf{x}_{-i}$ such that of all the types other than A, B has strictly the highest level of investment. Formally, $\mathbf{x}_{-i}$ is fixed such that $X_B(\mathbf{x}_{-A}) = \overline{X}_{-A}$ and there is no type $C \in \Theta$ such that $C \neq B$ and $X_C(\mathbf{x}_{-A}) = \overline{X}_{-A}$. Furthermore, $\mathbf{x}_{-i}$ is fixed such that $X_A(\mathbf{x}_{-i}) = X_B - 1$. In other words, given $\mathbf{x}_{-i}$, i's investment decision is pivotal between the status quo final type profile and a final type profile in which all players are of type B. We have that

$$\Delta_{A,B}^{retain} := E\left[\pi_A(x_i = 1, \mathbf{x}_{-i})\right] - E\left[\pi_A(x_i = 0, \mathbf{x}_{-i})\right] + c \tag{3.35}$$

where $\mathbf{x}_{-i}$ is such that $X_A(\mathbf{x}_{-i}) = \overline{X}_{-A}(\mathbf{x}_{-i}) - 1$. The quantity $\Delta_{A,B}^{retain}$ is thus the individual marginal benefit from investing, neglecting the cost of investment c, to each investing type-A player when they invest as much as type-B players and type-B players are the most-investing other type. Given other players' investments, i needs to invest to retain her own type.

Note that if $\Delta_A^{convert} > c$, then the best response of i (recalling that she is of type A, by assumption) to $\mathbf{x}_{-i}$ is to invest, and type-A players will have a total investment level of $\overline{X}_{-A} + 1$.[79] If $\Delta_A^{convert} \leq c$ and there exists some type B such that $\Delta_{A,B}^{retain} > c$ then X_A is the total investment level in the mutual best response by type B. Otherwise, the mutual best response of type-B players is that none of them invest.

Proposition 3.2 *At most two types invest in any pure equilibrium.*

Proof: Suppose *a contrario* that three or more types have non-zero investment. Then either three types have total investment equal to $\overline{X}(\mathbf{x}^*)$, or fewer than three do, in which case by hypothesis there is some type θ for which $X_\theta(\mathbf{x}^*) > 0$ and $X_\theta(\mathbf{x}^*) < \overline{X}(\mathbf{x}^*)$. In the former case, all investing players have a strictly profitable unilateral deviation to not investing, since

[79] Here I assume players do not invest in the case of indifference.

the final type profile would be unaffected were they to do so. In the latter case, all investing players of type θ have a strictly profitable unilateral deviation to not investing. ■

The intuition for Proposition 3.2 is straightforward: the conversion technology means that at most one type can successfully 'attack', and if two types are successfully 'defending', any further 'defence' is redundant.

To gain further insight into the incentives involved in ideological games in generality, fix an equilibrium investment profile $\mathbf{x}^*$ and again consider $\Delta_A^{convert}$. Intuitively, this is the marginal benefit from investing for a type-A player who has the "casting vote" between converting all other types and retaining the initial type profile. We have the following decomposition

$$\begin{aligned} \Delta_A^{convert} &= \overbrace{[w_A(\mathbf{z}^*(\mathbf{A});\mathbf{A}) - w_A(\mathbf{z}^*(\mathbf{A});\theta^{(\mathbf{1})})]}^{ideological\ incentive\ to\ convert} \\ &\quad + \underbrace{[w_A(\mathbf{z}^*(\mathbf{A});\theta^{(\mathbf{1})}) - w_A(\mathbf{z}^*(\theta^{(\mathbf{1})});\theta^{(\mathbf{1})})]}_{instrumental\ incentive\ to\ convert} \\ &= \Delta_A^{conv\,(id)} + \Delta_A^{conv\,(in)} \end{aligned} \tag{3.36}$$

where the ideological incentive $\Delta_A^{conv\,(id)} = w_A(\mathbf{z}^*(\mathbf{A});\mathbf{A}) - w_A(\mathbf{z}^*(\mathbf{A});\theta^{(\mathbf{1})})$ and the instrumental incentive $\Delta_A^{conv\,(in)} = w_A(\mathbf{z}^*(\mathbf{A});\theta^{(\mathbf{1})}) - w_A(\mathbf{z}^*(\theta^{(\mathbf{1})});\theta^{(\mathbf{1})})$. By (3.26), the ideological incentive $\Delta_A^{conv\,(id)} = 0$ if A is pragmatic. To understand this decomposition, first consider i's expected increase in meta-utility if the final type profile remained the same as the initial one (i.e. $\theta^{(\mathbf{1})}$) but there were an off-equilibrium path deviation from this final-stage outcome to the outcome $\mathbf{z}^*(\mathbf{A})$. This expected increase in i's meta-utility is the "instrumental" term in the decomposition. Second, now taking this off-equilibrium path deviation as a starting point consider a shift in the final type profile from $\theta^{(\mathbf{1})}$ to $\mathbf{A}$, i.e. to the final type profile that induces outcome $\mathbf{z}^*(\mathbf{A})$ in equilibrium. The resulting increase in meta-utility – attributable to the shift in final types, as opposed to actions – is the ideological incentive in the decomposition. In a similar fashion, $\Delta_{A,B}^{retain}$ can be decomposed as follows.

$$\begin{aligned}\Delta_{A,B}^{retain} &= \overbrace{[w_A(\mathbf{z}^*(\theta^{(\mathbf{1})});\theta^{(\mathbf{1})}) - w_A(\mathbf{z}^*(\mathbf{B});\theta^{(\mathbf{1})})]}^{instrumental\ incentive\ to\ retain} \\ &\quad + \underbrace{[w_A(\mathbf{z}^*(\mathbf{B});\theta^{(\mathbf{1})}) - w_A(\mathbf{z}^*(\mathbf{B});\mathbf{B})]}_{ideological\ incentive\ to\ retain} \\ &= \Delta_{A,B}^{ret\,(id)} + \Delta_{A,B}^{ret\,(in)}\end{aligned} \tag{3.37}$$

The decomposition of $\Delta_{A,B}^{retain}$ can be understood as follows. Starting this time with the ideological incentive, consider i's expected increase in meta-utility in moving from the final type profile $\mathbf{B}$ (which will occur if she does not invest) to the status quo of $\theta^{(\mathbf{1})}$, but holding constant the action profile as that induced by $\mathbf{B}$. This term captures the increase in meta-utility for i if we consider the direct impact of $\mathbf{B}$ only and neglect the indirect effect via the resultant switch in the outcome of the game of life Γ. From this off-equilibrium path starting point, we can then consider a subsequent shift in the action profile, which is picked up by the instrumental incentive.

So far I have considered ideological games in a very wide-ranging setting, to provide a general theory of the notions of ideology and pragmatism. To investigate the way ideology and pragmatism determine equilibrium investment behaviour, I will focus on the case of two players of different types, i.e. $N = \{1, 2\}$ and $\Theta = \{A, B\}$, and set the cost of investment $c = 0$.[80] I restrict attention to this setting partly because doing so shows the core differences between pragmatism and ideology in terms of players' investment incentives, without needing to consider potentially complicated mixed equilibria. Additionally, in a two-player setting, the conversion technology is a natural assumption with no compelling alternative specification: a player is converted iff her opponent invests and she does not.[81] In this sense, the results derived are intuitively appealing. In the case of several players of several types, in contrast, there do appear to be natural alternative specifications to the conversion technology considered.[82]

[80] For simplicity, I assume players do not invest, given other players' investments, in the case of indifference.

[81] An alternative would be that a player is converted with some interior probability iff her opponent invests and she does not, but clearly this would not change the analysis in a material way.

[82] One such alternative technology would be that a player is converted to a given type with a probability given by a Tullock contest function.

Different specifications would drive different results.

To arrive at a sufficient condition for one player to convert the other, we need to make a refinement of the concept of strong ideology. A player of initial type $\theta \in \Theta$ and the associated meta-payoffs $w_\theta(.)$ are *perfectly ideological* if and only if (i) they are strongly ideological and (ii) the following inequality holds.

$$\mathbf{z}^*(\theta, \theta_{-i}) = \arg\max_{\mathbf{z} \in Z^n} w_\theta(\mathbf{z}; \theta, \theta_{-i}) \tag{3.38}$$

where θ_{-i} is the profile of other players' types in which all players are of type θ.[83] (3.38) says that if all players are of type θ, the unique Nash equilibrium $\mathbf{z}^*(\theta, \theta_{-i})$ is the highest-ranked outcome of Γ for type θ. We then have the following result.

Proposition 3.3 *Suppose two players of different types play an ideological game where $c = 0$. (1) If both types are (weakly or strongly) ideological, then there exists an equilibrium in which both invest. (2) If both types are strongly ideological, equilibrium requires that both invest. (3) If one type is perfectly ideological and the other is pragmatic, only the perfectly ideological player will invest.*

Proof: Let A and B be the two types. To prove Proposition 3.3(1), first suppose A and B are ideological. By definition, the ideological payoffs $w_A(.)$ and $w_B(.)$ each satisfy (3.24), which implies that $\Delta_{A,B}^{retain} > 0$ and $\Delta_{B,A}^{retain} > 0$. Suppose the type-$A$ player invests. Then the other player's best response is to invest, as $\Delta_{B,A}^{retain} > 0$. This is an equilibrium investment profile, as $\Delta_{A,B}^{retain} > 0$. Turning to Proposition 3.3(2), now suppose that in addition to satisfying (3.24), $w_A(.)$ and $w_B(.)$ each satisfy (3.25), implying that $\Delta_A^{convert} > 0$ and $\Delta_B^{convert} > 0$. As $\Delta_A^{convert} > 0$ and $\Delta_{A,B}^{retain} > 0$, investing is strictly dominant for the type-A player; as $\Delta_B^{convert} > 0$ and $\Delta_{B,A}^{retain} > 0$, investing is also strictly dominant for the type-B player. Hence the unique equilibrium is the profile in which both players invest in the first stage and play $\mathbf{z}^*(A, B)$ in Γ. Finally, to prove Proposition 3.3(1), now suppose

[83] The strongly ideological devout type in section 3.2, for example, is perfectly ideological, since if all players possess strongly devout preferences, all abstain, which is the best outcome of the game of life Γ as judged by that preference type.

instead that A is perfectly ideological and B is pragmatic. In this case, since strong ideology is a necessary condition of perfect ideology, investing is strictly dominant for the type-A player. For the type-B (pragmatic) player, investing is optimal iff $w_B(\mathbf{z}^*(A,B);(A,B)) > w_B(\mathbf{z}^*(A,A);(A,A)) = w_A(\mathbf{z}^*(A,A);(A,A))$, where the last inequality follows from the condition in (3.29). However, (3.38) implies that $\mathbf{z}^*(A,A) = \arg\max_{\mathbf{z}\in Z^2} w_A(\mathbf{z};(A,A))$, while (3.25) implies that $(A,A) = \arg\max_{\theta\in\Theta^2} w_A(\mathbf{z};(A,A))$ for any $\mathbf{z}\in Z$. As $w_A(\mathbf{z}^*(A,A);(A,A))$ is therefore the highest payoff $w_A(.)$, by (3.27), $w_B(\mathbf{z}^*(A,A);(A,A))$ is the highest payoff in $w_B(.)$ and so $w_B(\mathbf{z}^*(A,B);(A,B)) \leq w_B(\mathbf{z}^*(A,A);(A,A))$. Hence the type-$B$ player does not invest in equilibrium. ■

Proposition 3.3(1) says that for two players of different ideological types, there is an equilibrium in which both invest. The reason for this is that an ideology requires its holder to want to retain her type at all costs; the ideological instrumental incentive to retain one's type outweighs any instrumental benefit from being converted. Consequently, even if neither player has a positive incentive to convert (which can be the case for players of weak, but not strong, ideologies) there exists an (inefficient) equilibrium in which both invest. Proposition 3.3(2) is simple to interpret: strong ideologies involve a preference to convert others and to defend one's own type, which outweighs any instrumental considerations. Consequently, investing is strictly dominant for a strongly ideological player whenever facing an opponent of a different type. Proposition 3.3(3) considers a perfectly ideological player against a pragmatic player. For a perfectly ideological player, ideological and instrumental incentives to convert are perfectly aligned. When facing an opponent of the same (perfect) ideology, the resulting equilibrium in Γ is the best possible outcome for the two players. This harmonious outcome is the reason why a pragmatic player finds it optimal to be converted in such a setting. Pragmatists "look through" their present identity in the sense that they evaluate ex-ante on the basis of their future preferences in Γ only, and are entirely indifferent as to other players' types in their own right.[84]

[84]In general, two players of different pragmatic types may result in any (pure) equilibrium investment profile, depending on the specification of the pragmatic types. The only general constraints regarding the

One remaining question I will consider in generality is how ideology can affect welfare, in the sense of ex-ante (meta-) payoffs summed over all players. Formally, define the *ex-ante efficiency* $E(\mathbf{x})$ of an investment profile as follows.

$$E(\mathbf{x}) := \sum_{i \in N} \left[w_i(\mathbf{z}^*(\theta^{(\mathbf{2})}(\mathbf{x})); \theta^{(\mathbf{2})}(\mathbf{x}))) - cx_i \right] - \sum_{i \in N} w_i(\mathbf{z}^*(\theta^{(\mathbf{1})}); \theta^{(\mathbf{1})})) \tag{3.39}$$

where, as before, $\theta^{(\mathbf{2})}(\mathbf{x})$ is the final type profile induced by investment profile $\mathbf{x}$ according to the conversion technology $f(.)$, and $c \geq 0$ is the cost of investment. Ex-ante efficiency is the difference between the first sum on the right hand side of (3.39), which gives the total ex-ante payoffs among all players when their investment choices are given by $\mathbf{x}$, and the second term, which is the sum of players' ex-ante payoffs when their investment choices are constrained to be zero (i.e. investment is not allowed). We then have the following result.

Lemma 3.4 *Fix an arbitrary type profile. (1) If precisely one type invests in equilibrium, then* $E(\mathbf{x}^*)$ *may be positive, negative or zero. (2) If two types invest in a pure strategy equilibrium then* $E(\mathbf{x}^*) < 0$.

Proof: To prove Lemma 3.4(1), I first construct the case where $E(\mathbf{x}^*) > 0$. Let there be two players, one of type A and one of type B, where $\Delta_A^{convert} > c$ and $\Delta_{B,A}^{retain} = 0$. In equilibrium, the player of type A invests and the other player does not, in which case the type-A player is strictly better off and the other player is indifferent, according to their ex-ante meta-preferences. Hence $E(\mathbf{x}^*) > 0$. The case where $E(\mathbf{x}^*) = 0$ occurs if for instance $\Delta_A^{convert} > c$ and $\Delta_{B,A}^{retain} = -\Delta_A^{convert}$. The case that $E(\mathbf{x}^*) < 0$ occurs if for instance $\Delta_A^{convert} = c' < c$ and

incentives to invest for a pair of players of different pragmatic types is that (i) if both have an incentive to convert, only one can have an incentive to retain, and (ii) if both have an incentive to retain, only one can have an incentive to convert. Each constraint is established by similar reasoning. Taking the example of the latter constraint, consider a type profile (A, B) where A and B are pragmatic types. Suppose both players invest in an equilibrium investment profile $\mathbf{x}^*$. Equilibrium requires $\Delta_{A,B}^{retain} \geq 0$ which implies $w_A(\mathbf{z}^*(\theta^{(\mathbf{1})}); \theta^{(\mathbf{1})}) \geq w_A(\mathbf{z}^*(\mathbf{B}); \mathbf{B})$. By Lemma 3.2, $\Delta_{A,B}^{retain} \geq 0$ iff $u_A(\mathbf{z}^*(\theta^{(\mathbf{1})})) > u_B(\mathbf{z}^*(\mathbf{B}))$. Likewise $\Delta_{B,A}^{retain} \geq 0$ iff $u_B(\mathbf{z}^*(\theta^{(\mathbf{1})})) > u_A(\mathbf{z}^*(\mathbf{A}))$. Hence $u_A(\mathbf{z}^*(\mathbf{A})) > u_A(\mathbf{z}^*(\theta^{(\mathbf{1})}))$ iff $u_B(\mathbf{z}^*(\theta^{(\mathbf{1})})) > u_B(\mathbf{z}^*(\mathbf{B}))$. But in this case, $\Delta_A^{convert} \geq 0$ implies $\Delta_B^{convert} < 0$. If both types have an incentive to convert and player 1 has an incentive to retain, then $(x_1, x_2) = (1, 0)$. If instead both have an incentive to convert and player 2 has an incentive to retain, then $(x_1, x_2) = (0, 1)$. If both players have an incentive to retain and one has an incentive to convert, then $(x_1, x_2) = (1, 1)$. The two constraints do not rule out there being a pair of pragmatic types in which neither has an incentive to retain nor to convert, in which case $(x_1, x_2) = (0, 0)$.

$\Delta_{B,A}^{retain} \in (0, c')$. Turning to Lemma 3.4(2), consider the general case of $n > 1$ players with an arbitrary initial type profile $\theta^{(\mathbf{1})} \in \Theta^n$, where Θ is an arbitrary set of types. If two types invest in any equilibrium $\mathbf{x}^*$ then $\theta^{(\mathbf{1})} = \theta^{(\mathbf{2})}$; otherwise, investing would not be optimal for players of at least one type, as their investment would not be pivotal. As the status quo is preserved, $E(\mathbf{x}^*) = -\sum_{i \in N} cx_i < 0$. ■

One point to note is that in cases where all types with support in the set of players are pragmatic, it can happen that in any equilibrium there is negative efficiency, just as for type distributions where ideological types are present. In the case of two players, one of pragmatic type A and the other of pragmatic type B, this can happen if for instance player A's receives a better outcome in Γ by playing a same-type player, i.e. $w_A(\mathbf{z}^*(A, A); (A, A)) > w_A(\mathbf{z}^*(A, B); (A, B))$, while player B receives the best outcome in Γ from playing an other-type player, i.e. $w_B(\mathbf{z}^*(A, B); (A, B) > w_B(\mathbf{z}^*(A, A); (A, A)) = w_A(\mathbf{z}^*(A, A); (A, A))$.

If just two players of different types are present, we can adapt the results of Proposition 3.3 to elicit a clearer difference in terms of efficiency between pragmatic and ideological types, as set out in Theorem 3.1.

Theorem 3.1 *Suppose two players of different types play an ideological game. (1) If both types are (weakly or strongly) ideological, then there exists a $c' > 0$ such that for any cost of investment $c \in (0, c')$, there exists an equilibrium $\mathbf{x}^*$ such that $E(\mathbf{x}^*) < 0$. (2) If both types are strongly ideological, then there exists a $c' > 0$ such that for any cost of investment $c \in (0, c')$, there exists a unique equilibrium $\mathbf{x}^*$ such that $E(\mathbf{x}^*) < 0$. (3) If one type is perfectly ideological and the other is pragmatic, there exists a $c' > 0$ such that for any cost of investment $c \in (0, c')$, there exists a unique equilibrium $\mathbf{x}^*$ such that $E(\mathbf{x}^*) > 0$.*

Proof: See Appendix C.1. The results in each case arise from the fact that for an arbitrarily small non-zero cost of investment, equilibrium investment profiles are those of Proposition 3.3. Because investment is nonetheless costly, if both types invest – as in Theorem 3.1(1) and 3.1(2) – then the final type profile is unchanged but both players are worse off than if they had not invested, as proved in Lemma 3.4. Theorem 3.1(3) simply follows from the fact

that for an arbitrarily small non-zero cost of investment, the perfectly ideological player finds investing strictly dominant and achieves her best payoff, while pragmatic player also achieves her best (meta-)payoff as a result of being converted to the perfect ideology.

3.4 Conclusion

In this chapter I defined the class of ideological games. In such games, a preference type is a set of first-order preferences over the outcomes of a game Γ, together with a set of meta-preferences over all players' first-order preferences. Rational players may bear costs to influence the type profile for two reasons. The first reason is instrumental: players may be able to achieve better outcomes in Γ if they change their opponent's type. The second is "ideological": players may have direct preferences over type profiles.

I provided a taxonomy of types as being "strongly ideological", "weakly ideological" or "pragmatic" and related these concepts to the notion of homophily (Proposition 3.1). A weakly ideological player always prefers to retain her own type, regardless of any instrumental benefits that may result from a better equilibrium outcome in Γ. This conception is based on an insight by Sen (1977). Strong ideology is a refinement of weak ideology: a strongly ideological player also prefers to change other players' types, regardless of any changes in payoffs in Γ she might receive. A pragmatic player, in contrast, is indifferent between all type profiles and is willing to have her type changed if, according to her new type, she would prefer the resulting equilibrium in Γ to that which would result if her type remained unchanged. An equilibrium investment profile contains at most two investing types (Proposition 3.2). I studied a measure of efficiency according to which an investment profile has positive efficiency if it improves aggregate welfare over the no-invest profile. In any two-player ideological game in which players have different types, if both types are ideological there is an equilibrium investment profile with negative efficiency (Theorem 3.1), since ideological players find it optimal to invest to retain their own type. If at least one type is strongly ideological, and the cost of investment is low enough, then the negative efficiency equilibrium is the only equilibrium. Finally, I defined "perfectly

ideological" types as strongly ideological meta-payoffs which, if held by all players, result in the best outcome of Γ as judged by that type. In any pairing of a perfectly ideological player with a pragmatic player, if the cost of investment is low enough to induce the former to invest, there is a unique equilibrium with positive efficiency.

3.5 References

1. Alonso, R. and O. Camara, 2016. Bayesian Persuasion with Heterogenous Priors. *Journal of Economic Theory*, vol. 165, pp. 672–70.

2. Bakshy, E., S. Messing and L. Adamic, 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*, vol. 348 no. 6239, pp. 1130-1132.

3. Becker, G. S., 1976. The Economic Approach to Human Behaviour. Chicago: University of Chicago Press.

4. Becker, G. S., 1981. A Treatise on the Family. Cambridge (Massachusetts): Harvard University Press.

5. Becker, G. S. and K. Murphy, 1988. A Theory of Rational Addiction. *Journal of Political Economy*, vol. 96(4), pp. 675-700.

6. Bénabou, R. J. M., 2008. Ideology. *Journal of the European Economic Association*, vol. 6, pp. 321–352.

7. Bénabou, R. J. M. and J. Tirole, 2004. Willpower and Personal Rules. *Journal of Political Economy*, vol. 112(4), pp. 848-886.

8. Binmore, K., 2009. Interpersonal Comparison of Utility. In D. Ross and H. Kincaid (eds), *The Oxford Handbook of Philosophy of Economics*. Oxford University Press.

9. Butler, D. M., Volden, C., Dynes, A. M. and B. Shor, 2017. Ideology, Learning, and Policy Diffusion: Experimental Evidence. *American Journal of Political Science*, vol. 61, pp. 37–49.

10. Cassels A., 1996. Ideology and International Relations in the Modern World. Routledge.

11. Currarini, S., M. O. Jackson and P. Pin, 2009. An Economic Model of Friendship: Homophily, Minorities, and Segregation. *Econometrica*, vol. 77(4), pp. 1003–1045.

12. Dawkins, R., 1976. The Selfish Gene. Oxford: Oxford University Press.

13. Dekel, E., B. Lipman and A. Rustichini, 2009. Temptation-Driven Preferences. *Review of Economic Studies*, vol. 76, pp. 937–971.

14. Dekel, E., J. Ely and O. Yilankaya, 2007. Evolution of Preferences. *Review of Economic Studies*, vol. 74, pp. 685–704.

15. Dennett, D., 1995. Darwin's Dangerous Idea: Evolution and the Meanings of Life. New York: Simon and Schuster.

16. Gul., F. and W. Pesendorfer, 2001. Temptation and Self-Control. *Econometrica*, vol. 69, pp. 1403-1435.

17. Harsanyi, J., 1955. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy*, vol. 63(4), pp 309-321.

18. Hunter, J. D., 1991. Culture Wars: The Struggle To Control The Family, Art, Education, Law, And Politics In America. New York: BasicBooks.

19. Khrushchev, N., 1959. On Peaceful Coexistence. *Foreign Affairs*, October.

20. Lerner, W., 1964. The Historical Origins of the Soviet Doctrine of Peaceful Coexistence. *Law and Contemporary Problems*, vol. 29(4), pp. 865-870.

21. Roemer, J. E., 1985. Rationalizing Revolutionary Ideology. *Econometrica*, vol. 53(1), pp. 85-108.

22. Segal, J. and A. Cover, 1989. Ideological Values and the Votes of U.S. Supreme Court Justices. *The American Political Science Review*, vol. 83(2), pp. 557-565.

23. Sen, A., 1977. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, vol. 6(4), pp. 317-344.

A Appendix to Chapter 1

A.1 Proofs

Proof of Lemma 1.3

Suppose there exists some equilibrium $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$ in which $a_i^\theta(\varepsilon) \in (0,1)$. At $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$, reciprocators must therefore be indifferent as to whether to cooperate, so on observing a reciprocator opponent, i's incentive to cooperate is $\triangle_i^c[r_\theta(\varepsilon)a_\theta^\theta(\varepsilon) + (1 - r_\theta(\varepsilon))a_\theta^0(\varepsilon)] = 0$. Let $\mathbf{a}'(\varepsilon)$ be the conditional action profile which is identical to $\mathbf{a}(\varepsilon)$ except that $a_\theta^\theta(\varepsilon) = 1$ (i.e. a randomly-selected reciprocator cooperates for sure on observing a reciprocator opponent). Then at $(\mathbf{r}(\varepsilon), \mathbf{a}'(\varepsilon))$, i's incentive to cooperate on observing a reciprocator opponent is $\triangle_i^c(1) > 0$, i.e. cooperation is strictly optimal. By symmetry, the same is also true for all other reciprocators. As the only one of i's three continuation values to depend on $a_\theta^\theta(\varepsilon)$ (on which it has positive dependence) is $v_i^\theta(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$, by (1.14) we have that $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}'(\varepsilon)) > \Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$ for all reciprocators. Let $\mathbf{r}'(\varepsilon)$ be a research profile such that $r_i'(\varepsilon) = \overline{p}$ iff $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}'(\varepsilon)) \geq 0$, which is a requirement of any equilibrium containing $\mathbf{a}'(\varepsilon)$. By symmetry, either $r_i'(\varepsilon) = \overline{p}$ for every reciprocator i or $r_i'(\varepsilon) = \underline{p}$ or every reciprocator i. It follows that $r_\theta'(\varepsilon) \geq r_\theta(\varepsilon)$, and hence at $(\mathbf{r}'(\varepsilon), \mathbf{a}'(\varepsilon))$, i's incentive to cooperate is on observing a reciprocator opponent $\triangle_i^c(a_\theta^\theta(\varepsilon)) = r_\theta'(\varepsilon)a_\theta^\theta(\varepsilon) + (1 - r_\theta'(\varepsilon))a_\theta^0(\varepsilon)) > 0$, and so $(\mathbf{r}'(\varepsilon), \mathbf{a}'(\varepsilon))$ meets equilibrium condition 2. By construction of $\mathbf{r}'(\varepsilon)$, equilibrium condition 1 is also met, so $(\mathbf{r}'(\varepsilon), \mathbf{a}'(\varepsilon))$ is an equilibrium. By Assumption 1.1, the former equilibrium, $(\mathbf{r}(\varepsilon), \mathbf{a}(\varepsilon))$, is not played. Analogous reasoning holds true in respect of equilibria in which $a_i^0(\varepsilon) \in (0,1)$. Lemma 1.2 established that materialists play d at every information set, implying $a_\theta^M(\varepsilon) = d$ in equilibrium. To complete the proof, it remains to rule out equilibria with $\mathbf{a}_\theta(\varepsilon) = dcd$. Suppose *a contrario* that such an equilibrium exists. Then for an arbitrary reciprocator i, $\triangle_i^c(r_\theta(\varepsilon)a_\theta^\theta(\varepsilon) + (1 - r_\theta(\varepsilon))a_\theta^0(\varepsilon)) < \triangle_i^c(\varepsilon r_\theta(\varepsilon)a_\theta^\theta(\varepsilon) + (1 - r_\theta(\varepsilon))a_\theta^0(\varepsilon)) + (1 - \varepsilon) \times 0)$, where the final zero comes from the fact that materialists never cooperate). But this inequality is false, as $\varepsilon \leq 1$. □

Proof of Lemma 1.4

Fix arbitrary $\varepsilon \in [0,1]$ and suppose *a contrario* that $r_\theta(\varepsilon) \in (\underline{p}, \overline{p})$, which implies that some non-zero measure of reciprocators play $r(\varepsilon) = \underline{p}$ and some non-zero measure of reciprocators play $r(\varepsilon) = \overline{p}$. Let i be some arbitrary reciprocator who plays $r(\varepsilon) = \overline{p}$ and fix θ. Then i's incentive to do research must be non-negative: $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) \geq 0$. By symmetry, this must also be true for any other reciprocator. Lemma 1.3 states there are three second-stage profiles to consider: $\mathbf{a}_\theta(\varepsilon) = ccd$, $\mathbf{a}_\theta(\varepsilon) = cdd$ and $\mathbf{a}_\theta(\varepsilon) = ddd$.

Suppose first that $\mathbf{a}_\theta(\varepsilon) = ccd$. In this case, $v(a_i^\theta, a_\theta^\theta) = v(a_i^\theta, a_\theta^0) = u(c,c)$ and $v(a_i^0, a_\theta^\theta) = v(a_i^0, a_\theta^0) = u(c,c)$, and so by (1.18), $\frac{\partial \Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))}{\partial r_\theta(\varepsilon)} = 0$, i.e. $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))$ is independent of $r_\theta(\varepsilon)$. As a result, if all reciprocators play $r(\varepsilon) = \overline{p}$, for each reciprocator i, $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) \geq 0$ as before, and so research profiles with $r_\theta(\varepsilon) = \overline{p}$ meet equilibrium condition 1. As $a_\theta^\theta(\varepsilon) = c = a_\theta^0(\varepsilon)$ by hypothesis, i's incentive to cooperate at each of her information sets is unaffected, so equilibrium condition 2 holds. Assumption 1.1 implies that the new research profile is played.

Second, suppose that $\mathbf{a}_\theta(\varepsilon) = cdd$. In this case, $\frac{\partial \Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon))}{\partial r_\theta(\varepsilon)} = \varepsilon \triangle p([u(c,c) - u(c,d)] - [u(d,c) - u(d,d)]) > 0$, which means that the incentive to do research for each reciprocator is strictly increasing in $r_\theta(\varepsilon)$. Again, if all reciprocators play $r(\varepsilon) = \overline{p}$, then for each reciprocator i, $\Delta_i^r(\varepsilon, \mathbf{r}_{-i}(\varepsilon), \mathbf{a}(\varepsilon)) \geq 0$ as before, and so research profiles with $r_\theta(\varepsilon) = \overline{p}$ meet equilibrium condition 1. As i's incentive to cooperate at all information sets is non-decreasing in $r_\theta(\varepsilon)$, $a_\theta^\theta(\varepsilon) = c$ still meets equilibrium condition 2. If $a_\theta^0(\varepsilon) = d$ no longer meets equilibrium condition 2, then ccd is played in which case $r_\theta(\varepsilon) = \overline{p}$, as argued above. If $a_\theta^0(\varepsilon) = d$ meets equilibrium condition 2, Assumption 1.1 implies that $r_\theta(\varepsilon) = \overline{p}$ is played; by the tie-breaking rule, every reciprocator chooses $r_i(\varepsilon) = \overline{p}$.

Third, and finally, suppose that $\mathbf{a}_\theta(\varepsilon) = ddd$. In this case, Lemma 1.1 implies that $r_i(\varepsilon) = \underline{p}$ for every reciprocator. $\square$

Proof of Proposition 1.1

To prove Proposition 1.1, I will first characterise the regions of parameter values for which each of the five candidate pure symmetric equilibria is played in the fully model, before applying the equilibrium selection criteria in Assumption 1.1 to obtain the unique equilibrium in question. The results in the case of the no-technology model then follow trivially.

To this end, it will be helpful to define $\triangle^{(r_i', \mathbf{a}_i')}(\varepsilon, r, \mathbf{a})$, the *incentive to deviate*:

$$\triangle^{(r_i', \mathbf{a}_i')}(\varepsilon, r, \mathbf{a}) := u_i(\varepsilon, r_i', \mathbf{r}_{-i}, \mathbf{a}_i', \mathbf{a}_{-i}) - u_i(r_i, \mathbf{r}_{-i}, \mathbf{a}_i, \mathbf{a}_{-i}) \tag{A.1}$$

where i is a reciprocator, $r \in \{\underline{p}, \overline{p}\}$ is a research choice, $\mathbf{a} \in [0,1]^3$ is a conditional action vector and $(\mathbf{r}_{-i}, \mathbf{a}_{-i})$ is a profile of (stage 1) research choices and (stage 2) conditional action vectors for all players other than i, in which all materialists play $(\underline{p}, ddd)$ and all reciprocators except i play $(r, \mathbf{a})$.[85] The quantity $\triangle^{(\underline{p}, ddd)}(\varepsilon, \overline{p}, ccd)$, for example, is the incentive for a reciprocator to deviate unilaterally from playing $(\overline{p}, cdd)$ to playing $(\underline{p}, ddd)$, when all other reciprocators play $(\overline{p}, cdd)$ (and all materialists play $(\underline{p}, ddd)$), given the share of reciprocators in the population is $\varepsilon \in [0,1]$. If and only if $\triangle^{(\underline{p}, ddd)}(\varepsilon, \overline{p}, ccd) > 0$, then i strictly prefers playing $(\underline{p}, ddd)$ to playing $(\overline{p}, cdd)$ in such a setting.

(i) Parameter values for which $(\underline{p}, ddd)$ ***is an equilibrium***

Recall that since (d, d) is a Nash equilibrium of Γ for all preference types, $(\underline{p}, ddd)$ is an equilibrium for all parameter values.

(ii) Parameter values for which $(\underline{p}, cdd)$ ***is an equilibrium***

Let the strategy played by all reciprocators be fixed as $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) = (\underline{p}, cdd)$. In the second stage, reciprocators always defect when meeting materialists and receive zero payoff from all such encounters. Non-zero payoffs only arise if a reciprocator is paired with another

[85] I continue to use the shorthand whereby conditional action vector $(1, 0, 1)$ is denoted *cdc*, with analogous notation for the other pure conditional action vectors.

reciprocator, which happens with probability $\varepsilon \in [0,1]$. Each reciprocator has expected utility $\varepsilon[\underline{p}^2(\theta b - 1) + \underline{p}(1-\underline{p})b - \underline{p}(1-\underline{p})]$. The first term within the square brackets is the product of the probability $\underline{p}^2$ which which a pair of reciprocators observe each other's type, and the payoff from mutual cooperation $(\theta - 1)b$. The second term is the probability that reciprocator i meets a reciprocator j who does not observes i's type while i observes j's, multiplied by the corresponding payoff $u(d,c) = b$. The third term is the probability $\underline{p}(1-\underline{p})$ that j observes i's type while i does not observe j's, multiplied by the corresponding payoff $u(c,d) = -1$. The incentives for a player to deviate unilaterally to one of the other pure strategy profiles (excluding $(\overline{p}, ddd)$, as this is strictly worse for any player than playing $(\underline{p}, ddd)$) are as follows.

- *Unilateral deviation to* $(\underline{p}, ddd)$ yields expected utility of $\varepsilon \underline{p} b$, as the deviating player meets a fellow reciprocator with probability ε, who will play c with probability $\underline{p}$ by hypothesis, yielding a payoff of $u(d,c) = b > 1$. The incentive for reciprocator i to deviate is $\triangle^{(\underline{p},ddd)}(\varepsilon, \underline{p}, cdd) = \varepsilon(-\underline{p}^2(\theta - 1)b + \underline{p})$, and so deviation is strictly profitable at any $\varepsilon > 0$ iff $(\theta - 1)b < \frac{1}{\underline{p}}$. This inequality is straightforward to interpret. The benefit to i of cooperating rather than defecting with an opponent known to be a reciprocator, net of the cost of cooperation, is $\underline{p}(\theta - 1)b$; if this does not exceed the cost of cooperation (of 1), then i will always find defection preferable.

- *Unilateral deviation to* $(\overline{p}, cdd)$ yields expected utility of $\varepsilon \underline{p}\overline{p}(\theta b - 1) + \varepsilon \underline{p}(1-\overline{p})b - \varepsilon\overline{p}(1-\underline{p}) - k$; note the inclusion of the fourth and final term, which is simply the cost of doing research. The incentive to deviate is $\triangle^{(\overline{p},cdd)}(\varepsilon, \underline{p}, cdd) = \varepsilon \triangle p(\underline{p}(\theta - 1)b) - 1) - k$, which means that, if $(\theta - 1)b \geq \frac{1}{\underline{p}}$, then deviation is strictly profitable iff $\varepsilon < \frac{k}{\triangle p[\underline{p}(\theta-1)b-1]}$. To understand this inequality, observe that if i is paired with another reciprocator j, then doing research increases the probability with which i observes j's type by amount $\triangle p$. The opponent j cooperates with probability $\underline{p}$, by hypothesis, and so i's expected utility increase resulting from increased mutual cooperation is $\triangle p[u_i(c,c) - u_i(d,c)] = \triangle p \underline{p}(\theta b - 1 - b)$. If j defects then i's expected utility increase is $\triangle p(1-\underline{p})[u_i(d,d) - u_i(c,d)] = (1-\underline{p})$; summing these two terms gives the denominator on the right hand

side of the inequality. If $(\theta-1)b < \frac{1}{\underline{p}}$, on the other hand, then deviation is strictly profitable at any $\varepsilon > 0$.

- *Unilateral deviation to* $(\underline{p}, ccd)$ yields expected utility of $\varepsilon\underline{p}(\theta b-1)-\varepsilon(1-\underline{p})-(1-\varepsilon)(1-\underline{p})$ and the incentive to deviate is $\triangle^{(\underline{p},ccd)}(\varepsilon,\underline{p},cdd) = (1-\underline{p})(1-\varepsilon\underline{p}(\theta b-1))$. Deviation is therefore profitable iff $\varepsilon \geq \frac{1}{\underline{p}(\theta-1)b}$. Intuitively, for deviation to be profitable, the expected benefit from increased mutual cooperation by cooperating with an opponent of unknown type net of the cost of cooperation, $\varepsilon\underline{p}(\theta-1)b$, must exceed the cost of cooperation (of 1).

- *Unilateral deviation to* $(\overline{p}, ccd)$ yields expected utility of $\varepsilon\underline{p}(\theta b-1)-\varepsilon(1-\underline{p})-(1-\varepsilon)(1-\overline{p})-k$ and the incentive to deviate is $\triangle^{(\overline{p},ccd)}(\varepsilon,\underline{p},cdd) = \varepsilon(\underline{p}(1-\underline{p})(\theta-1)b-\triangle p)-k-(1-\overline{p})$, which is clearly strictly negative for all $\varepsilon \in [0,1]$ if $(\theta-1)b \leq \frac{\triangle p}{\underline{p}(1-\underline{p})}$. If $(\theta-1)b > \frac{\triangle p}{\underline{p}(1-\underline{p})}$, then the incentive to deviate is strictly positive iff $\varepsilon > \frac{(1-\overline{p})+k}{(1-\underline{p})\underline{p}(\theta-1)b-\triangle p}$. By inspection, it is also less than the expected utility from deviating to $(\underline{p}, ccd)$ iff $\varepsilon > 1-\frac{k}{\triangle p}$.

Comparing these results, if $(\theta-1)b < \frac{1}{\underline{p}}$, then $(\underline{p}, cdd)$ is never an equilibrium because there is a profitable deviation to $(\underline{p}, ddd)$. If $(\theta-1)b \geq \frac{1}{\underline{p}}$ then it must also be the case that $(\theta-1)b > \frac{\triangle p}{\underline{p}(1-\underline{p})}$, since $\triangle p < (1-\underline{p})$. In this case, $\frac{(1-\overline{p})+k}{(1-\underline{p})\underline{p}(\theta-1)b-\triangle p} > \frac{1}{\underline{p}(\theta-1)b}$ iff $\frac{k}{\triangle p} > (1-\frac{1}{\underline{p}(\theta-1)b})$ and $\frac{k}{\underline{p}\triangle p(\theta-1)b-\triangle p} > \frac{1}{\underline{p}(\theta-1)b}$ iff $\frac{k}{\triangle p} > (1-\frac{1}{\underline{p}(\theta-1)b})$. Consequently, if $\frac{k}{\triangle p} \geq 1-\frac{1}{\underline{p}(\theta-1)b}$ then it is an equilibrium for $\varepsilon \leq \frac{1}{\underline{p}(\theta-1)b}$ and if $\frac{k}{\triangle p} < 1-\frac{1}{\underline{p}(\theta-1)b}$ then it is an equilibrium for $\varepsilon \leq \frac{k}{\triangle p[\underline{p}(\theta-1)b-1]}$. For fixed research cost k, if the impact of research $\triangle p$ is sufficiently small then even when $\varepsilon = 1$, reciprocators do not have a strong enough incentive to do research, if they are constrained to play $\mathbf{a}_\theta(\varepsilon) = cdd$.

(iii) Parameter values for which $(\overline{p}, cdd)$ is an equilibrium.

Suppose now that $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) = (\overline{p}, cdd)$. As in the previous case, non-zero payoffs only arise if a reciprocator is paired with another reciprocator, which happens with probability $\varepsilon \in [0,1]$.

Each reciprocator has expected utility $\varepsilon[\overline{p}^2(\theta b-1)+\overline{p}(1-\overline{p})b-\overline{p}(1-\overline{p})]-k$. The incentives for a reciprocator to deviate are as follows.

- *Unilateral deviation to* $(\underline{p}, ddd)$ yields expected utility of $\varepsilon\overline{p}b$. The incentive for reciprocator i to deviate is $\triangle^{(\underline{p},ddd)}(\varepsilon,\overline{p},cdd)=\varepsilon(-\overline{p}^2(\theta-1)b+\overline{p})+k$, and so if $(\theta-1)b\leq\frac{1}{\overline{p}}$, then deviation is strictly profitable for any any$\varepsilon\in[0,1]$ and $k>0$. If $(\theta-1)b>\frac{1}{\overline{p}}$, then deviation is strictly profitable iff $\varepsilon<\frac{k}{\overline{p}[\overline{p}(\theta-1)b-1]}$.

- *Unilateral deviation to* $(\underline{p}, cdd)$ yields expected utility of $\varepsilon\underline{p}\overline{p}(\theta b-1)+\varepsilon\overline{p}(1-\underline{p})b-\varepsilon\underline{p}(1-\overline{p})$. The incentive to deviate is $\triangle^{(\underline{p},cdd)}(\varepsilon,\overline{p},cdd)=\varepsilon\triangle p(-\overline{p}(\theta-1)b+1)+k$, which means that, if $(\theta-1)b\geq\frac{1}{\overline{p}}$, then deviation is strictly profitable iff $\varepsilon<\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}$. If $(\theta-1)b<\frac{1}{\overline{p}}$, on the other hand, then deviation is strictly profitable at any $\varepsilon>0$. Reciprocators only find research mutually optimal for a large enough population share ε because the value of undertaking research to reciprocator is in improving the chance of being able to cooperate if she meets a fellow reciprocator. The more other reciprocators there are, the stronger this incentive.

- *Unilateral deviation to* $(\underline{p}, ccd)$ yields expected utility of $\varepsilon\overline{p}(\theta b-1)-\varepsilon(1-\overline{p})-(1-\varepsilon)(1-\underline{p})$. The incentive to deviate is $\triangle^{(\underline{p},ccd)}(\varepsilon,\overline{p},cdd)=\varepsilon((1-\overline{p})\overline{p}(\theta b-1)+\triangle p)-(1-\underline{p})+k$. Deviation is therefore strictly profitable iff $\varepsilon>\frac{(1-\underline{p})-k}{(1-\overline{p})\overline{p}(\theta-1)b+\triangle p}$.

- *Unilateral deviation to* $(\overline{p}, ccd)$ yields expected utility of $\varepsilon\overline{p}(\theta b-1)-\varepsilon(1-\overline{p})-(1-\varepsilon)(1-\overline{p})-k$. The incentive to deviate is $\triangle^{(\overline{p},ccd)}(\varepsilon,\overline{p},cdd)=\varepsilon(1-\overline{p})\overline{p}(\theta-1)b-(1-\overline{p})$, which is strictly positive iff $\varepsilon>\frac{1}{\overline{p}(\theta-1)b}$.

Comparing these results, first, we see that $\frac{k}{\overline{p}[\overline{p}(\theta-1)b-1]}<\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}$ (for all allowable parameter values) and so if $(\theta-1)b>\frac{1}{\underline{p}}$, then whenever a deviation to $(\underline{p}, ddd)$ is profitable, one to $(\underline{p}, cdd)$ is also profitable. This is hardly surprising; the inequality $(\theta-1)b>\frac{1}{\underline{p}}$ ensures that cooperation with a known reciprocator is optimal, as discussed in case (ii) above. Another observation is that if $k>(1-\underline{p})$, it is always profitable to deviate to play $(\underline{p}, ccd)$, and so $(\overline{p}, cdd)$ is not an equilibrium.

Assume $(\theta-1)b > \frac{1}{\overline{p}}$ (otherwise there is always a profitable deviation to $(\underline{p}, ddd)$). Iff $\frac{k}{\triangle p} > 1 - \frac{1}{\overline{p}(\theta-1)b}$ then $\frac{1}{\overline{p}(\theta-1)b} < \frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}$, and so $(\overline{p}, cdd)$ is not an equilibrium at any $\varepsilon \in (0,1)$. Iff $\frac{k}{\triangle p} \leq 1 - \frac{1}{\overline{p}(\theta-1)b}$, then $\frac{1}{\overline{p}(\theta-1)b} < \frac{(1-\underline{p})-k}{(1-\overline{p})\overline{p}(\theta-1)b+\triangle p}$.[86] It follows that $(\overline{p}, cdd)$ is an equilibrium for $\varepsilon \in [\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}, \frac{1}{\overline{p}(\theta-1)b}]$; deviation to $(\underline{p}, cdd)$ is profitable below this interval because the reciprocator population is insufficient to make research generate enough increased cooperation, whereas deviation to $(\overline{p}, ccd)$ is worthwhile above this interval because there are enough reciprocators to make it optimal to cooperate with a partner of unknown type.

(iv) Parameter values for which $(\underline{p}, ccd)$ is an equilibrium.

Suppose that $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) = (\underline{p}, ccd)$. In this case, if two reciprocators meet always cooperate, regardless of whether they observe each other's type. Each reciprocator has expected utility $\varepsilon(\theta b - 1) - (1-\varepsilon)(1-\underline{p})$, where the first term is the payoff from mutual cooperation $(\theta b - 1)$ multiplied by the probability ε of being paired with a reciprocator and the second term is the probability of meeting a materialist without learning their type, multiplied by the payoff of -1 from outcome (d, c). The incentives for a reciprocator to deviate are as follows.

- *Unilateral deviation to* $(\underline{p}, ddd)$ yields expected utility of εb. The incentive for reciprocator i to deviate is $\triangle^{(\underline{p},ddd)}(\varepsilon, \underline{p}, ccd) = -\varepsilon((\theta-1)b - \underline{p}) + (1-\underline{p})$, which is strictly positive iff $\varepsilon < \frac{1-\underline{p}}{(\theta-1)b-\underline{p}}$.

- *Unilateral deviation to* $(\underline{p}, cdd)$ yields expected utility of $\varepsilon(\underline{p}(\theta b - 1) + (1-\underline{p})b)$. The incentive to deviate is $\triangle^{(\underline{p},cdd)}(\varepsilon, \underline{p}, ccd) = -\varepsilon(1-\underline{p})(\theta-1)b + (1-\underline{p})$, which is strictly positive iff $\varepsilon < \frac{1}{(\theta-1)b}$; the threshold arises from the fact that playing $a_i^0(\varepsilon) = c$ rather than $a_i^0(\varepsilon) = d$ has a gross cost of 1 for an expected gross benefit of $\varepsilon(\theta-1)b$. Note that

[86] If $(\theta-1)b > \frac{1}{\overline{p}}$ and $\frac{k}{\triangle p} = 1 - \frac{1}{\overline{p}(\theta-1)b}$, then $\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]} = \frac{(1-\underline{p})-k}{(1-\overline{p})\overline{p}(\theta-1)b+\triangle p} = \frac{1}{\overline{p}(\theta-1)b}$. Why this "triple point" occurs is as follows. First, if all reciprocators other than i play $(\overline{p}, cdd)$, then when $\varepsilon = \frac{1}{\overline{p}(\theta-1)b}$, i is indifferent between playing $a_i^0 = c$ and $a_i^0 = d$. Second, if the effective cost of research $\frac{k}{\triangle p}$ equals the materialist population share $(1-\varepsilon)$, then a player is indifferent between research and not when playing ccd, as paying k for research in such a situation has value solely in reducing the probability (by $\triangle p$) of being victim of free-riding to a materialist. Hence if $(1-\varepsilon) = \frac{k}{\triangle p} = 1 - \frac{1}{\overline{p}(\theta-1)b}$, then i is indifferent between playing $(\overline{p}, cdd)$ and deviating to $(\underline{p}, ccd)$ or $(\overline{p}, ccd)$. But then she is indifferent between cooperating or not whether or not she does research, which means deviating to $(\underline{p}, cdd)$ also leaves her indifferent.

$\triangle^{(\underline{p},cdd)}(\varepsilon,\underline{p},ccd) > \triangle^{(\underline{p},ddd)}(\varepsilon,\underline{p},ccd)$; i is better off playing $a_i^\theta(\varepsilon) = c$ than $a_i^\theta(\varepsilon) = c$ because other reciprocators always cooperate with her.

- *Unilateral deviation to* $(\overline{p},cdd)$ yields expected utility of $\varepsilon(\overline{p}(\theta b-1)+(1-\overline{p})b)-k$. The incentive to deviate is $\triangle^{(\overline{p},ccd)}(\varepsilon,\underline{p},ccd) = -\varepsilon(1-\overline{p})(\theta-1)b-\varepsilon\triangle p+(1-\underline{p})-k$, which is strictly positive iff $\varepsilon < \frac{(1-\underline{p})-k}{(1-\overline{p})(\theta-1)b+\triangle p}$.

- *Unilateral deviation to* $(\overline{p},ccd)$ yields expected utility of $\varepsilon(\theta b-1)-(1-\varepsilon)(1-\overline{p})-k$. The incentive to deviate is $\triangle^{(\overline{p},ccd)}(\varepsilon,\overline{p},cdd) = (1-\varepsilon)\triangle p-k$, which is strictly positive iff $\varepsilon < 1-\frac{k}{\triangle p}$. The inequality has a simple interpretation. The value of research to a reciprocator if other reciprocators play $a_\theta^0(\varepsilon) = c$ is in reducing the probability of being suckered when meeting a materialist (an encounter that happens with probability $(1-\varepsilon)$), to yield a net expected benefit of $\triangle p[u_i(c,d)-u_i(d,d)] = \triangle p$. Consequently, if the cost of research k does not exceed $(1-\varepsilon)\triangle p$, then research is optimal, independently of other reciprocators' research choices. For a fixed cost of research $k > 0$, the threshold below which research is optimal increases in the impact of research $\triangle p$. Likewise, for fixed impact of research, the threshold decreases in the cost of research.

For all the candidate unilateral deviations, deviation is profitable below a certain threshold. This is an intuitive result; if we consider an all-reciprocator population (i.e. $\varepsilon = 1$), then $(r_\theta(\varepsilon),\mathbf{a}_\theta(\varepsilon)) = (\underline{p},cdd)$ is clearly an equilibrium; research has no value to reciprocators as there is no risk of suffering a free-riding opponent, and mutual cooperation will happen with certainty. We have that $(\underline{p},ccd)$ is an equilibrium for $\varepsilon \geq \max\{\frac{(1-\underline{p})-k}{(1-\overline{p})(\theta-1)b+\triangle p}, 1-\frac{k}{\triangle p}, \frac{1}{(\theta-1)b}\}$. It turns out that $\frac{(1-\underline{p})-k}{(1-\overline{p})(\theta-1)b+\triangle p} < \frac{1}{(\theta-1)b}$ iff $\frac{k}{\triangle p} > 1-\frac{1}{(\theta-1)b}$, which is true iff $\frac{1-\underline{p}-k}{(1-\overline{p})(\theta-1)b+\triangle p} > 1-\frac{k}{\triangle p}$.[87] Consequently, if $\frac{k}{\triangle p} < 1-\frac{1}{(\theta-1)b}$, then $(\underline{p},ccd)$ is an equilibrium for $\varepsilon \geq 1-\frac{k}{\triangle p}$. If

[87] If $\frac{k}{\triangle p} = 1-\frac{1}{(\theta-1)b}$, then $\frac{(1-\underline{p})-k}{(1-\overline{p})(\theta-1)b+\triangle p} = 1-\frac{k}{\triangle p} = \frac{1}{(\theta-1)b}$. We can understand why using reasoning analogous to that in footnote 86. If all reciprocators other than i play $(\underline{p},ccd)$, then when the effective cost of research $\frac{k}{\triangle p}$ equals the materialist population share $(1-\varepsilon)$, a player is indifferent between doing research and not when playing ccd, as paying k for research in such a situation has value solely in reducing the probability (by $\triangle p$) of being victim of free-riding to a materialist. Also, if $\varepsilon = \frac{1}{(\theta-1)b}$, then i is indifferent between cooperating and not when she does not learn her opponent's type. But then she is indifferent between cooperating or not whether or not she does research, which means deviating to $(\overline{p},cdd)$ also leaves her indifferent.

on the other hand $\frac{k}{\triangle p} > 1 - \frac{1}{(\theta-1)b}$, then $(\underline{p}, ccd)$ is an equilibrium for $\varepsilon \geq \frac{1}{(\theta-1)b}$.

(v) Parameter values for which $(\overline{p}, ccd)$ is an equilibrium.

Finally, suppose that $(r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)) = (\overline{p}, ccd)$. Each reciprocator has expected utility $\varepsilon(\theta b - 1) - (1-\varepsilon)(1-\overline{p}) - k$. The incentives for a reciprocator to deviate are as follows.

- *Unilateral deviation to* $(\underline{p}, ddd)$ yields expected utility of εb. The incentive for reciprocator i to deviate is $\triangle^{(\underline{p},ddd)}(\varepsilon, \overline{p}, ccd) = -\varepsilon((\theta-1)b - \overline{p}) + (1-\overline{p}) + k$, which is strictly positive iff $\varepsilon < \frac{(1-\overline{p})+k}{(\theta-1)b-\overline{p}}$.

- *Unilateral deviation to* $(\underline{p}, cdd)$ yields expected utility of $\varepsilon(\underline{p}(\theta b - 1) + (1-\underline{p})b) > \varepsilon b$. The incentive to deviate is $\triangle^{(\underline{p},cdd)}(\varepsilon, \overline{p}, ccd) = -\varepsilon(1-\underline{p})(\theta-1)b + \varepsilon\triangle p + (1-\overline{p}) + k$, which is strictly positive iff $\varepsilon < \frac{(1-\overline{p})+k}{(1-\underline{p})(\theta-1)b-\triangle p}$.

- *Unilateral deviation to* $(\overline{p}, cdd)$ yields expected utility of $\varepsilon(\overline{p}(\theta b - 1) + (1-\overline{p})b) - k$. The incentive to deviate is $\triangle^{(\overline{p},cdd)}(\varepsilon, \overline{p}, ccd) = -\varepsilon(1-\overline{p})(\theta-1)b + (1-\overline{p})$, which is strictly positive iff $\varepsilon < \frac{1}{(\theta-1)b}$, since for population shares below this threshold the expected gross benefit of cooperation $\varepsilon(\theta-1)b$ for a reciprocator who doesn't learn her opponent's type is less than 1, the gross cost of cooperation.

- *Unilateral deviation to* $(\underline{p}, ccd)$ yields expected utility of $\varepsilon(\theta b - 1) - (1-\varepsilon)(1-\overline{p})$. The incentive to deviate is $\triangle^{(\underline{p},ccd)}(\varepsilon, \overline{p}, ccd) = -(1-\varepsilon)\triangle p + k$, which is strictly positive iff $\varepsilon > 1 - \frac{k}{\triangle p}$. As the incentive to do research when playing *ccd* is to reduce the risk of cooperating when one's opponent free-rides, the threshold is independent of other reciprocators' research choices.

Comparing the incentives to deviate, we see that $(\overline{p}, ccd)$ is an equilibrium if both $\varepsilon \geq \max\{\frac{(1-\overline{p})+k}{(1-\underline{p})(\theta-1)b}, \frac{1}{(\theta-1)b}\}$ and $\varepsilon < 1 - \frac{k}{\triangle p}$. If $\frac{k}{\triangle p} > 1$, then the second of these conditions cannot be met. If $\frac{k}{\triangle p} \in (1, \frac{1}{(\theta-1)b})$, then there is no value of $\varepsilon \in [0,1]$ for which $(\overline{p}, ccd)$ is an equilibrium, since it is always better not to do research when playing *ccd* in stage 2. If, on the other hand, $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$, then $(\overline{p}, ccd)$ is an equilibrium for $\varepsilon \in (\frac{1}{(\theta-1)b}, 1 - \frac{k}{\triangle p})$.

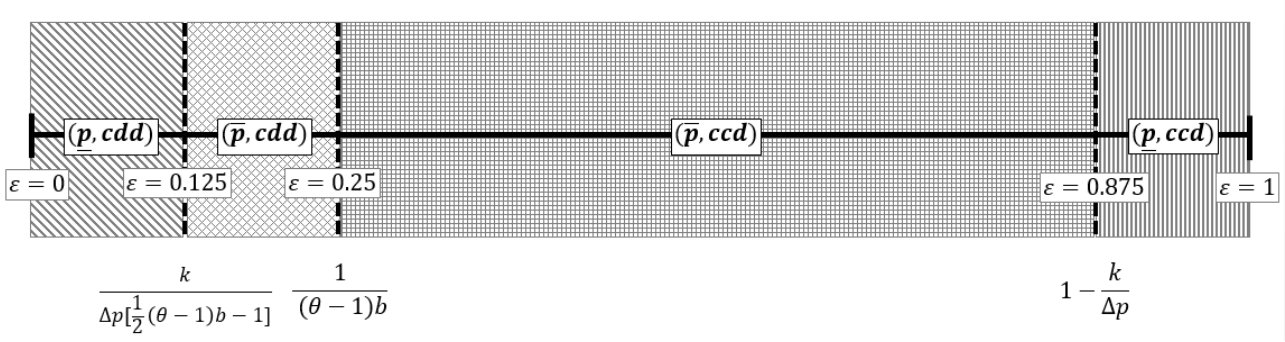
Having characterised all the symmetric pure strategy equilibria in the two-stage game, it remains to apply the equilibrium selection criteria to arrive at a unique equilibrium. Fixing arbitrary ε, Assumption 1.1 implies that if more than one profile is played, then whichever has the highest number in the following list is selected: (1) $(\underline{p}, ddd)$; (2) $(\underline{p}, cdd)$; (3) $(\overline{p}, cdd)$; (4) $(\underline{p}, ccd)$; (5) $(\overline{p}, ccd)$. This can be seen as follows. Comparing $(\overline{p}, ccd)$ with $(\underline{p}, ccd)$, the only value of ε at which both strategy profiles can be played is $1 - \frac{k}{\triangle p}$, at which reciprocators are (independently of each other's research choices) indifferent between doing research and not; the tie-breaking rule of Assumption 1.1 implies that $(\overline{p}, ccd)$ is played. Comparing $(\underline{p}, ccd)$ with $(\overline{p}, cdd)$, $\varepsilon(\underline{p} + (1 - \underline{p})) = \varepsilon$ is the *ex ante* probability i's opponent cooperates with her in the former profile, which (provided $\varepsilon > 0$) is strictly greater than $\frac{1}{2}\varepsilon$, the probability her opponent cooperates in the latter. Likewise, $(\overline{p}, ccd)$ yields an opponent's *ex ante* cooperation probability of ε. The profile $(\underline{p}, ccd)$ yields an opponent's *ex ante* cooperation probability of $\underline{p}\varepsilon < \frac{1}{2}\varepsilon$. Finally, $(\underline{p}, ddd)$ clearly induces zero cooperation.

As $(\overline{p}, ccd)$ is the highest-ranked profile, if $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$, it is played at $\varepsilon \in [\frac{1}{(\theta-1)b}, 1 - \frac{k}{\triangle p}]$; otherwise, it is not played. As the next-highest ranked profile, $(\underline{p}, ccd)$ is played at $\varepsilon \geq \frac{1}{(\theta-1)b}$ unless $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$ in which case it is played at $\varepsilon \in (1 - \frac{k}{\triangle p}, 1]$. The third-ranked profile, $(\overline{p}, cdd)$, satisfies the two equilibrium conditions (setting aside Assumption 1.1) at $\varepsilon \in [\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}, \frac{1}{\overline{p}(\theta-1)b}]$ if$\frac{k}{\triangle p} \leq 1 - \frac{1}{\overline{p}(\theta-1)b}$; otherwise, it is not played in equilibrium. As either $(\overline{p}, ccd)$ or $(\underline{p}, ccd)$ is played at $\varepsilon \geq \frac{1}{(\theta-1)b}$, Assumption 1.1 implies that $(\overline{p}, cdd)$ can only be played in $\varepsilon \in [\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}, \frac{1}{(\theta-1)b}]$. The interval is non-null only if $\frac{k}{\triangle p} \leq \overline{p} - \frac{1}{(\theta-1)b}$, so $(\overline{p}, cdd)$ is played at $\varepsilon \in [\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}, \frac{1}{(\theta-1)b}]$ iff $\frac{k}{\triangle p} \leq \overline{p} - \frac{1}{(\theta-1)b}$. The profile $(\underline{p}, cdd)$ is played at $\varepsilon < \frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}$ if $\frac{k}{\triangle p} \leq \overline{p} - \frac{1}{(\theta-1)b}$ and for $\varepsilon < \frac{1}{(\theta-1)b}$ if $\frac{k}{\triangle p} > \overline{p} - \frac{1}{(\theta-1)b}$.

It remains to show that if research is done in equilibrium, it is done iff $\varepsilon \in [\varepsilon_1, \varepsilon_2]$, where $0 < \varepsilon_1 \leq \varepsilon_2 < 1$. If $\frac{k}{\triangle p} \geq 1 - \frac{1}{\overline{p}(\theta-1)b}$, then research is not done if $\varepsilon < \frac{1}{(\theta-1)b}$, while if $\varepsilon \geq \frac{1}{(\theta-1)b}$, Proposition 1.1(1) implies that research is either not done or it done iff $\varepsilon \in [\frac{1}{(\theta-1)b}, 1 - \frac{k}{\triangle p}]$. If $\frac{k}{\triangle p} < 1 - \frac{1}{\overline{p}(\theta-1)b}$, research is done iff $\varepsilon \in [\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}, 1 - \frac{k}{\triangle p}] = [\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}, \frac{1}{(\theta-1)b}) \cup [\frac{1}{(\theta-1)b}, 1 - \frac{k}{\triangle p}]$. This concludes the proof of Proposition 1.1. ■

Figure 18 illustrates an example equilibrium, where $\underline{p} = 0.3$, $\overline{p} = 0.5$, $b = 4$, $\theta = 2$ and $k = 0.025$. As $\frac{k}{\triangle p} = 0.125 < \frac{5}{6} = 1 - \frac{1}{\underline{p}(\theta-1)b}$, by Proposition 1.1(2), reciprocators play $\mathbf{a}_\theta(\varepsilon) = (\overline{p}, cdd)$ for $\varepsilon \in [\frac{k}{\triangle p[\overline{p}(\theta-1)b-1]}, \frac{1}{(\theta-1)b}) = [0.125, 0.25)$ while by Proposition 1.1(1), they play $\mathbf{a}_\theta(\varepsilon) = (\overline{p}, cdd)$ for $\varepsilon \in [\frac{1}{(\theta-1)b}, 1 - \frac{k}{\triangle p}] = [0.25, 0.875]$.

Figure 18: Equilibrium played when $\theta = 2$, $\underline{p} = 0.3$, $\overline{p} = 0.5$, $b = 4$ and $k = 0.025$



Remark. *Reciprocators do research for population shares below the free-riding threshold only if they also do research at the threshold (where they play ccd). This is because for cdd to be played at all in equilibrium, it is necessary that* $(\theta - 1)b \geq \frac{1}{\overline{p}}$*; otherwise, there is a profitable deviation to* $(\underline{p}, ddd)$*. This implies that the free-riding threshold* $\frac{1}{(\theta-1)b} < \frac{1}{\overline{p}}$*. Now, if all reciprocators play* $\mathbf{a}_\theta(\varepsilon) = cdd$ *for some* $\varepsilon < \frac{1}{(\theta-1)b}$*, each finds research optimal if* $k \leq \frac{1}{\overline{p}}\varepsilon\triangle p(\theta - 1)b < \overline{p}\triangle p$*. At the threshold, if all reciprocators play ccd, then research is optimal for each one if* $k \leq \triangle p(1 - \frac{1}{(\theta-1)b})$*, where* $\triangle p(1 - \frac{1}{(\theta-1)b}) > \overline{p}\triangle p$*.*

Proof of Lemma 1.5

Fix arbitrary $\overline{p} \in (0, 1)$ and $\underline{p} \in (0, \overline{p})$. The quantities $\triangle F(\theta, \varepsilon, \underline{p}, cdd)$, $\triangle F(\theta, \varepsilon, \overline{p}, cdd)$, $\triangle F(\theta, \varepsilon, \underline{p}, ccd)$ and $\triangle F(\theta, \varepsilon, \overline{p}, ccd)$ are all strictly increasing in $\varepsilon > 0$. As a result, for given k, if a single pure strategy is played for $\varepsilon \in [\varepsilon_1, \varepsilon_2] \subseteq [0, 1]$, then

$$\overline{\varepsilon}(\theta) \equiv \sup_{\varepsilon \in (0,1]} \{\{0\} \cup \{\varepsilon : \forall \hat{\varepsilon} \in (0, \varepsilon), \triangle F[\hat{\varepsilon}, r_\theta(\hat{\varepsilon}), \mathbf{a}_\theta(\hat{\varepsilon})] > 0\}\} \notin (\varepsilon_1, \varepsilon_2) \tag{A.2}$$

This property clearly holds if $\triangle F[\varepsilon_1, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)] \leq 0$, in which case, $\overline{\varepsilon}(\theta) \leq \varepsilon_1$. If on the

other hand $\triangle F[\varepsilon_1, r_\theta(\varepsilon), \mathbf{a}_\theta(\varepsilon)] > 0$, then either (i) $\forall \hat{\varepsilon} \in (0, \varepsilon_1)$, $\triangle F[\hat{\varepsilon}, r_\theta(\hat{\varepsilon}), \mathbf{a}_\theta(\hat{\varepsilon})] > 0$ or (ii) $\exists \hat{\varepsilon} \in (0, \varepsilon_1) : \triangle F[\hat{\varepsilon}, r_\theta(\hat{\varepsilon}), \mathbf{a}_\theta(\hat{\varepsilon})] \leq 0$. In case (i),

$$\sup_{\varepsilon \in (0,1]} \{\{0\} \cup \{\varepsilon : \forall \hat{\varepsilon} \in (0, \varepsilon) , \triangle F[\hat{\varepsilon}, r_\theta(\hat{\varepsilon}), \mathbf{a}_\theta(\hat{\varepsilon})] > 0\}\} \geq \varepsilon_2 \tag{A.3}$$

while in case (ii),

$$\sup_{\varepsilon \in (0,1]} \{\{0\} \cup \{\varepsilon : \forall \hat{\varepsilon} \in (0, \varepsilon) , \triangle F[\hat{\varepsilon}, r_\theta(\hat{\varepsilon}), \mathbf{a}_\theta(\hat{\varepsilon})] > 0\}\} < \varepsilon_1 \tag{A.4}$$

Likewise, if a pure strategy is played for $\varepsilon \in (\varepsilon_1, \varepsilon_2] \subseteq [0, 1]$, then $\overline{\varepsilon}(\theta) \notin (\varepsilon_1, \varepsilon_2)$. The attainable share $\overline{\varepsilon}(\theta)$ can therefore be obtained by calculating the relative fitness at each infimum of the interval of population shares at which a symmetric pure strategy profile is played, and listing those at which relative fitness is strictly greater than zero. For example, suppose there are three symmetric strategy profiles played, with infima of 0, $\varepsilon_1 > 0$ and $\varepsilon_2 > \varepsilon_1$, such that relative fitness at $\varepsilon = 0$ and $\varepsilon = \varepsilon_2$ is strictly positive, but that at ε_1 is negative. The value of $\overline{\varepsilon}(\theta)$ is then the supremum of the interval such it and all lower intervals of symmetric pure strategy profiles have a infimum on the list. In the example, it is the supremum of the interval whose infimum is zero, i.e. $\overline{\varepsilon}(\theta) = \varepsilon_1$. Using this technique, I will prove each part of Lemma 1.5 in turn.

To analyse the version of the model without the discovery technology, let us impose the constraint $r = \underline{p}$ on all players in the first stage, so that only the second stage of the game involves strategic play. Accordingly, equilibrium condition 2 is imposed with research choice set to $r = \underline{p}$ for all players. Lemma 1.3, which derives from equilibrium condition 2 and the incentive to cooperate, still goes through, i.e. reciprocators all play either *ddd*, *cdd* or *ccd*. In establishing whether any one of these profiles is an equilibrium it is sufficient to check whether an arbitrary reciprocator has a profitable unilateral deviation to either of the other two profiles. For arbitrary $\varepsilon \in [0, 1]$, *ccd* is therefore an equilibrium if either $(\underline{p}, ccd)$ or $(\overline{p}, ccd)$ is an equilibrium in the full model; similarly, *cdd* is an equilibrium if either $(\underline{p}, cdd)$ or is an equilibrium in the full model. By Proposition 1.1, $\mathbf{a}_\theta(\varepsilon) = ccd$ iff $\varepsilon \geq \frac{1}{(\theta-1)b}$, while for

$\varepsilon < \frac{1}{(\theta-1)b}$, $\mathbf{a}_\theta(\varepsilon) = cdd$ if $(\theta-1)b \geq \frac{1}{\underline{p}}$, which is true by assumption, since $\Theta(b, \underline{p}) = [1+\frac{1}{b\underline{p}}, \infty)$. We have $\triangle F[\theta, \varepsilon, (\underline{p}, cdd)] > 0$ for $\varepsilon > 0$, and so $\overline{\varepsilon}_{no\,tech}(\theta) \geq \frac{1}{(\theta-1)b}$. Checking relative fitness at the lower bound of the region where $\mathbf{a}_\theta(\varepsilon) = ccd$ is played, i.e. at $\varepsilon = \frac{1}{(\theta-1)b}$, yields $\triangle F[\theta, \frac{1}{(\theta-1)b}, (\underline{p}, ccd)] = \frac{\underline{p}(b-1)}{(\theta-1)b} - (1-\underline{p})$. Hence $\triangle F[\theta, \frac{1}{(\theta-1)b}, (\underline{p}, ccd)] > 0$ iff $\theta < 1 + \frac{\underline{p}(b-1)}{b(1-\underline{p})}$, and so $\overline{\varepsilon}_{no\,tech}(\theta) = 1$ iff $\theta \leq 1 + \frac{\underline{p}(b-1)}{b(1-\underline{p})}$; otherwise, $\overline{\varepsilon}_{no\,tech}(\theta) = \frac{1}{(\theta-1)b}$. □

Proof of Theorem 1.1

Theorem 1.1 characterises $\overline{\varepsilon}(\theta)$ in the the full model, when technology is present and where, by hypothesis, $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$. The proof uses the technique of the proof of Lemma 1.5, i.e. calculating the relative fitness at each infimum of the interval of population shares at which a symmetric pure strategy profile is played.

Setting $\overline{p} = 1$, Proposition 1.1 implies that $(\overline{p}, cdd)$ is played for $\varepsilon \in [\frac{k}{\triangle p[(\theta-1)b-1]}, \frac{1}{(\theta-1)b})$ and $(\overline{p}, ccd)$ is played for $\varepsilon \in [\frac{1}{(\theta-1)b}, 1 - \frac{k}{\triangle p}]$. At the lower bound of the region where $(\overline{p}, cdd)$ is played, $\triangle F[\theta, \frac{k}{\triangle p[(\theta-1)b-1]}, (\overline{p}, cdd)] = \frac{k(b-1)}{\triangle p[(\theta-1)b-1]} - k$. Hence $\triangle F[\theta, \frac{k}{\triangle p[(\theta-1)b-1]}, (\overline{p}, cdd)] > 0$ iff $\theta < 1 + \frac{1}{b} + \frac{b-1}{(1-\underline{p})b}$.

I start by proving Theorem 1.1(1). It can easily be verified that that $1+\frac{1}{b}+\frac{b-1}{(1-\underline{p})b} \leq 1+\frac{1}{b\underline{p}}$ iff $\underline{p}(b+1) - \underline{p}^2 - 1 \leq 0$, which implies that if $\underline{p}(b+1) - \underline{p}^2 \leq 0$, $\overline{\varepsilon}(\theta) = \frac{k}{\triangle p[(\theta-1)b-1]} < \frac{1}{(\theta-1)b}$ for any $\theta \in \Theta(b) = [1 + \frac{1}{b\underline{p}}, \infty)$. Suppose instead that $\underline{p}(b+1) - \underline{p}^2 - 1 > 0$ and define $\theta' := 1 + \frac{1}{b} + \frac{b-1}{(1-\underline{p})b}$. In this case, if $\theta \geq \theta'$, then $\triangle F(\theta, \frac{k}{\triangle p[(\theta-1)b-1]}, \overline{p}, cdd) \leq 0$ and so $\overline{\varepsilon}(\theta) = \frac{k}{\triangle p[(\theta-1)b-1]} < \frac{1}{(\theta-1)b}$. Consequently, for $\theta' = 1 + \frac{1}{b} + \frac{b-1}{(1-\underline{p})b}$, if $\underline{p}(b+1) - \underline{p}^2 - 1 \leq 0$ or $\theta \geq \theta'$, then $\overline{\varepsilon}(\theta) = \frac{k}{\triangle p[(\theta-1)b-1]} < \overline{\varepsilon}_{no\,tech}(\theta)$. By inspection, as $k \to 0$, $\overline{\varepsilon}(\theta) \to 0$, as required.

The proof of Theorem 1.1(2) is then as follows. If $\theta < \theta'$, then relative fitness at the lower bound where $(\overline{p}, cdd)$ is played is positive, i.e. $\triangle F(\theta, \frac{k}{\triangle p[(\theta-1)b-1]}, \overline{p}, cdd) > 0$, and so $\overline{\varepsilon}(\theta) \geq \frac{1}{(\theta-1)b}$. For any $\varepsilon \in [0, 1]$, we have that $\triangle F(\theta, \varepsilon, \overline{p}, ccd) - \triangle F(\theta, \varepsilon, \underline{p}, ccd) = \varepsilon \triangle p(b-1) + \triangle p - k$. As $\frac{k}{\triangle p} \leq 1 - \frac{1}{(\theta-1)b}$, $\triangle F(\theta, \varepsilon, \overline{p}, ccd) > \triangle F(\theta, \varepsilon, \underline{p}, ccd)$. Setting $\varepsilon = \frac{1}{(\theta-1)b}$, if $\triangle F(\theta, \frac{1}{(\theta-1)b}, \underline{p}, ccd) > 0$ then $\triangle F(\theta, \frac{1}{(\theta-1)b}, \overline{p}, ccd) > 0$, and so $\overline{\varepsilon}(\theta) \geq \overline{\varepsilon}_{no\,tech}(\theta)$, as required.■

A.2 Numerical examples with an imperfect discovery technology

Theorem 1.1 assumes a perfect discovery technology, i.e. $\overline{p} = 1$. However, for $\overline{p} < 1$ it is in general also possible to find parameter values for which a discovery technology raises or lowers the attainable share. To illustrate this point, Figures 19 and 20 below provide examples in which $\overline{\varepsilon}(\theta) < \overline{\varepsilon}_{no\,tech}(\theta)$ for a (lower) type and $\overline{\varepsilon}(\theta) > \overline{\varepsilon}_{no\,tech}(\theta)$ another (higher) type, respectively.

In Figure 19, the discovery technology is used by reciprocators while they still blindly defect. The boost in relative fitness they receive is represented by the difference between the lower end of the solid black line at $\varepsilon = 0.25$ and the lower end of the vertical dotted line beneath it.

Figure 19: Relative fitness by reciprocator population share in equilibrium and for selected non-equilibrium profiles when $k = 0.1$, $\underline{p} = 0.25$, $\overline{p} = 0.5$, $b = 6$ and $\theta = 1\frac{2}{3}$

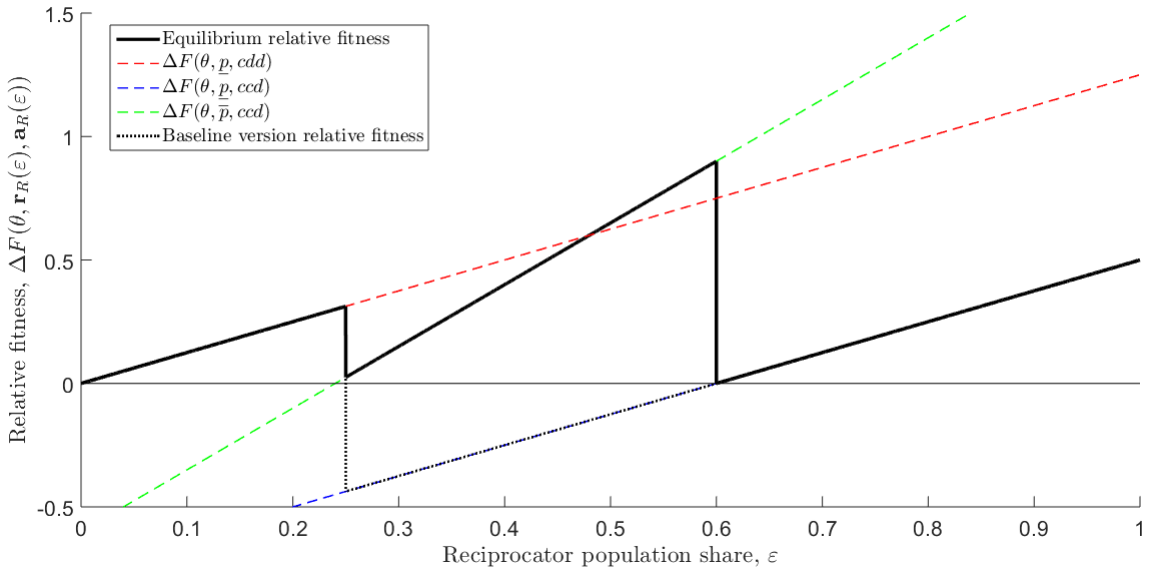
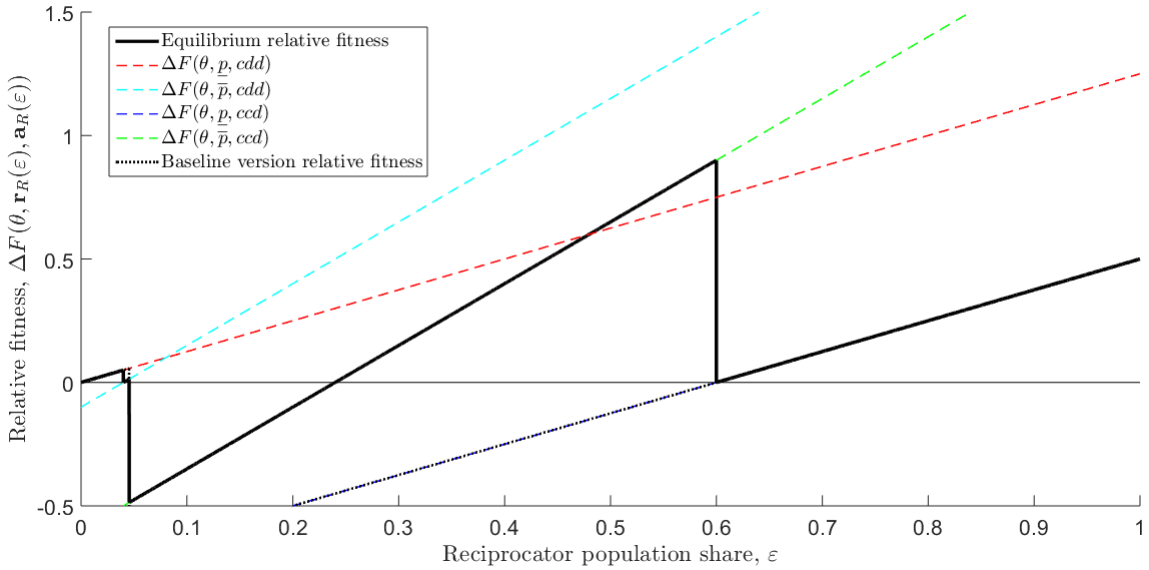


Figure 20 illustrates Theorem 1.1(2), for the case $\theta = 4.7$. In this case, the reciprocator type values mutual cooperation enough to do research below the free-riding threshold (which is at $\varepsilon = \frac{1}{22}$); at the lower bound of the interval of population shares where research is done in equilibrium, reciprocators are too small a share of the population to have positive relative

fitness. Importantly, it is only below the free-riding threshold, for low population shares where reciprocators do blind defection, that doing research can reduce relative fitness for reciprocators, because it is in this region that they have an incentive to 'overpay' for research, to promote mutual cooperation (as opposed to having the incentive to avoid being free-ridden, which is the case at higher reciprocator population shares).[88]

Figure 20: Relative fitness by reciprocator population share in equilibrium and for selected non-equilibrium profiles when $k = 0.1$, $\underline{p} = 0.25$, $b = 6$ and $\theta = 4\frac{2}{3}$



[88] In figure 20, the choice of parameters makes the difference $\overline{\varepsilon}(\theta) - \overline{\varepsilon}_{no\,tech}(\theta)$ fairly small. If I had made the definition of attainable share weaker, replacing the requirement for strictly positive fitness in (1.20) with one for weakly positive fitness, this would remove the constraint that $(\theta - 1)b \geq \frac{1}{\underline{p}}$ for $\overline{\varepsilon}(\theta) \neq \overline{\varepsilon}_{no\,tech}(\theta)$, widening the ranges of parameter values for which $\overline{\varepsilon}(\theta) < \overline{\varepsilon}_{no\,tech}(\theta)$ for some types and $\overline{\varepsilon}(\theta) > \overline{\varepsilon}_{no\,tech}(\theta)$ for others. However, the amended definition of attainable share would be less compelling than the one I have adopted, as it would not longer capture the population share to which an arbitrarily small initial share of reciprocators would grow under payoff-monotone fitness dynamics.

B Appendix to Chapter 2

B.1 Proofs

Proof of Theorem 2.1

Existence. The proof for existence is standard: note that $x_i = 0$ for all $i \in N$ is always an equilibrium. Note also that from (2.6), for any $\mathcal{Q}_\gamma$, the (finite) strategy profile containing $x_i = 1$ and $x_j = 0$ is always an equilibrium, for all $i \in \mathcal{Q}_\gamma$ and $j \notin \mathcal{Q}_\gamma$.

Uniqueness. The uniqueness of the ME can be proved by contradiction, as follows. Suppose there exist two distinct ME, $\mathbf{x}$ and $\mathbf{x}'$ with $\mathbf{x} \neq \mathbf{x}'$. Since both $\mathbf{x}$ and $\mathbf{x}'$ are ME, there must exist at least one i and one j such that $x_i < x'_i$ and $x_j > x'_j$. Let us now construct an equilibrium $\hat{\mathbf{x}}$, whereby $\hat{x}_i = \max\{x_i, x'_i, x^*_i\}$ for all i, where x^*_i is i's best-response to $\hat{\mathbf{x}}_{-i}$, and with $x^*_i \geq \max\{x_i, x'_i\}$. To see why $\hat{\mathbf{x}}$ is an equilibrium, note first that since neighbours' actions are weak local complements, and as i's neighbours play 1 in $\hat{\mathbf{x}}$ if they play it either in $\mathbf{x}$ or in $\mathbf{x}'$, then i's best response must be to play 1 if it is her action either in $\mathbf{x}$ or in $\mathbf{x}'$. Second, if i plays 0 in both those profiles, then we simply have that $\hat{x}_i = x^*_i \in \{0, 1\}$. Thus $\hat{\mathbf{x}}$ is an equilibrium. Clearly, $\hat{x}_i \geq x_i$, with the inequality strict for at least one j. The same is true for x'_i, contradicting the claim that $\mathbf{x}$ and $\mathbf{x}'$ are both ME. That the result holds for any parameter values is straightforward given the definition of γ and $\mathcal{S}_\gamma$. □

Characterisation. It remains to prove that $\mathbf{x}^*$, whereby $x^*_i = 1$ iff $i \in \mathcal{Q}_\gamma$, is the ME. We proceed by contradiction. Suppose that there is another equilibrium $\mathbf{x}'$ that has a higher contribution level than $\mathbf{x}^*$. Denote by $\mathcal{Q}' \supseteq \mathcal{Q}_\gamma$ the set of players who contribute in $\mathbf{x}'$. Then, there must exist at least one $j \notin \mathcal{Q}_\gamma$ and $j \in \mathcal{Q}'$ with $x'_j = 1$ and $x^*_j = 0$, implying the following claim.

Claim B.1 *There must exist one player $k \in \mathcal{Q}'$, with $x^*_k = 1$, with*

$$\Delta\left(\mathbf{x}'_{j \in N_k(G)}\right) = -\gamma + \Psi\left(x_i, c'_i, d'_i\right) < 0 \tag{B.1}$$

Proof: Suppose not. Then, this entails that for all $i \in \mathcal{Q}'$:

$$\Psi\left(x_i, c_i, d_i\right) > \gamma \tag{B.2}$$

But by definition of $\mathcal{Q}_\gamma$, expression (B.2) entails that $i \in Q' \Rightarrow i \in \mathcal{Q}_\gamma$. But this contradicts the hypothesis that $j \notin \mathcal{Q}_\gamma$ and $j \in \mathcal{Q}'$. □

Note finally that Claim B.1 entails that $\mathbf{x}'$ is not an equilibrium: indeed player k can strictly increase her payoffs by switching to $x = 0$. This completes the proof. ■

Proof of Proposition 2.1

We first show that the contribution level at the ME decreases in γ. Note that in any G and for any θ, an increase in γ always (weakly) reduces $\mid \mathcal{Q}_\gamma(G) \mid$, by construction. We know from Theorem 2.1 that a decrease in $\mid \mathcal{Q}_\gamma(G) \mid$ yields a decrease in the contribution level.

Second, we provide a sketch proof of Proposition 2.1(1)-2.1(3). Let $i,\ j \in \mathcal{Q}_\gamma(G)$ be two unlinked players and consider adding a link between them so that the resulting network is $G + ij$. We can check whether i and j continue to contribute in the ME. Suppose that $x_i^* = x_j^* = 1$. Given these fixed actions, for any player $k \neq i, j$, $s_k^\gamma(G + ij) = s_k^\gamma(G)$ and so $k \in \mathcal{Q}_\gamma(G + ij)$ if and only if $k \in \mathcal{Q}_\gamma(G)$. Further, we have $s_i^\gamma(G + ij) = s_i^\gamma(G) + 1$ and $s_j^\gamma(G + ij) = s_j^\gamma(G) + 1$. By Assumption 2.1(4), $\Psi\left(1, s_i^\gamma, (k_i - s_i^\gamma)\right)$ is increasing in s_j^γ, and so $i,\ j \in \mathcal{Q}_\gamma(G)$ by the characterisation via inequality (2.6). Hence the ME is unchanged. Proposition **2.1**(2) can be established by similar reasoning: adding a link between two players outside $\mathcal{Q}_\gamma(G)$ weakly expands $\mathcal{Q}_\gamma(G)$, and so does the deletion of a link (e.g. between a conditional cooperator and a materialist). In both cases, total contributions thus weakly increase. Lastly, the example provided with Figure 13 in the main text proves Proposition **2.1**(3). This completes the proof. ■

Proof of Proposition 2.2

Consider first the minimal number of j's neighbours changing their action necessary for j to change her action following i's change of type, when $\theta_i = \theta_R$ initially:[89]

$$r_j^*(\theta, \gamma, G) := \min_{r \in \mathbb{N}_+} \left\{ r : \Psi\left(x_i, (c_j^* - r,), (d_j^* + r)\right) \leq \gamma \right\} \tag{B.3}$$

We first show that if $r_j^* > 0$ and j is connected to at least r_j^* i-susceptible players, then j is i-susceptible.[90] Suppose *a contrario* that j is not i-susceptible. Given $r_j^* > 0$, this means that $x_j^* = x_j' = 1$ by definition, where x_j^* (x_j') is j's action before (after) i's type change.

Denote by r_j the number of i-susceptible players j is connected to, with $r_j \geq r_j^*$ by hypothesis. Since $x_j^* = x_j = 1$, then for j's action to be optimal we require

$$r_j < \min_{r \in \mathbb{N}_+} \left\{ r : \Psi\left(x_i, (c_j^* - r,), (d_j^* + r)\right) \leq \gamma \right\} \tag{B.4}$$

But it then follows that $r_j < r_j^*(\theta, \gamma, G)$, which contradicts the hypothesis that $r_j \geq r_j^*$. □

Second, we show that if j is i-susceptible, then j is connected to at least r_j^* i-susceptible players. Suppose *a contrario* that j is i-susceptible but is connected to $r_j < r_j^*$ i-susceptible players. Since j is i-susceptible, then by definition $x_j^* = 1 \neq x_j' = 0$, where x_j^* (x_j') is j's action before (after) i's type change.

Since $x_j^* = 1 \neq x_j' = 0$, then it follows that $r_j \geq r_j^*(\theta, \gamma, G)$, which contradicts the hypothesis that $r_j < r_j^*$. □

These two statements together complete the proof. ■

[89] In the other case, that $\theta_i = \theta_M$ initially, susceptible players will be those that switch from $x_j^* = 0$ to $x_j' = 1$. In this other case, there is simply a sign change in the definition of $r_j^* \in \mathbf{r}^*$, which is now given by $r_j^*(\theta, \gamma, G) := \min_{r \in \mathbb{N}_+} \left\{ r : \Psi\left(x_j^*, (c_j^* + r,), (d_j^* - r)\right) \geq \gamma \right\}$. The proof for this case is analogous and thus omitted.

[90] Note that if $r_j^* = 0$, then j is already free-riding, and thus can never be influenced by i's change of type and is thus "unsusceptible".

Proof of Proposition 2.3

We begin with Proposition 2.3(1). Again we assume that $\theta_i = \theta_R$ initially; the opposite case is analogous, and the proof is thus omitted. Recall first that if $j \in \mathcal{S}^i(G)$, then $x_j^* = 1$, and thus $j \in \mathcal{Q}_\gamma$. Then, observe that adding a link between a player j and a player k, with $j, k \in \mathcal{S}^i(G)$, increases r_j^* and r_k^*, which weakly reduces i's and k's susceptibility to i. Hence $\mathcal{S}^i(G + jk) \subseteq S^i(G)$, where $G + jk$ denotes the addition of a link between j and k to G. Second, consider removing a link between a player j and a player k, with $j, k \in \mathcal{S}^i(G)$. Either (i) $j, k \in \mathcal{S}^i(G - jk)$, or (ii) $j \notin \mathcal{S}^i(G - jk)$ or $k \notin \mathcal{S}^i(G - jk)$, but not both, or (iii) $j, k \notin \mathcal{S}^i(G - jk)$. In case (i), clearly $\mathcal{S}^i(G - jk) = \mathcal{S}^i(G)$; we will show that in cases (ii) and (iii), $\mathcal{S}^i(G - jk) \subset S^i(G)$.

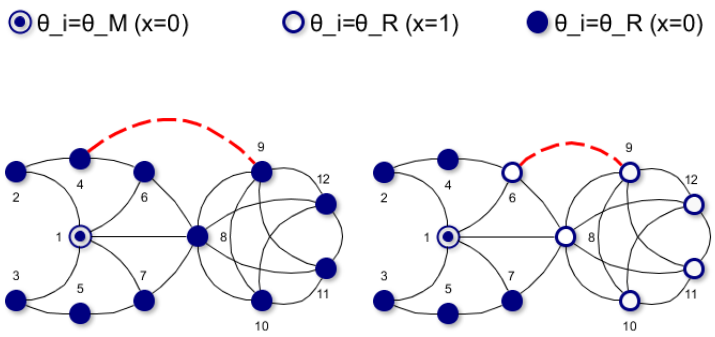
In case (ii), first, suppose without loss of generality that $j \notin \mathcal{S}^i(G - jk)$. In this case, either (ii.a) j contributes before and after i's switch in network $G - ij$, or (ii.b) j free-rides before and after i's switch in network $G - ij$ (which happens if $j \notin \mathcal{Q}_\gamma(G - jk)$). In case (ii.a), every other player has less incentive to switch action following i's switch than in network G, and so $\mathcal{S}^i(G - jk) \subset S^i(G)$. In case (ii.b), after i's switch in either network, j and k free-ride, and so there is the same ME after i's switch in both G and $G - jk$, and thus $\mathcal{S}^i(G - jk) \subset S^i(G)$. Second, in case (iii), if j and k play the same action as each other, then the reasoning for case (ii) directly applies. It remains to consider what happens if one player (say j) contributes before and after i's switch in network $G - ij$ while the other (say k) free-rides before and after i's switch. First, note that from Proposition 2.1, $\mathcal{Q}_\gamma(G - jk) \subset \mathcal{Q}_\gamma(G)$. Second, note that every player $\ell \in \mathcal{Q}_\gamma(G - jk)$ has weakly less incentive to free-ride after i's switch than they did in network G (since j now contributes after i's switch). Hence $\mathcal{S}^i(G - jk) \subset S^i(G)$.This completes the proof to Proposition 2.3(1).

Proposition 2.3(2) follows from the following observation: if the ME does not change following the addition or deletion of a link between two players $j, k \notin \mathcal{S}^i(G)$, than it follows that r_j^* is left unchanged for all $j \in \mathcal{S}^i(G)$, and the number of links to i-susceptible players for any $j \in \mathcal{S}^i(G)$ does not change either. Hence, i's influence cannot decrease. However, the

deletion of a link between two contributing conditional cooperators j and k outside $\mathcal{S}^i(G)$ strictly decreases r_j^* and r_k^*, which can make either or both i-susceptible, which increases i's influence. Second, the addition of a link between a contributing conditional cooperator j and a materialist k, for example, also strictly decreases r_j^*, which can make j i-susceptible. Hence, the addition or the deletion of a link between two players $j, k \notin \mathcal{S}^i(G)$ either increases $\mathcal{S}^i(G)$ or leaves it unchanged when the ME does not change. The example in the main text shows that if the ME changes, however, than $\mathcal{S}^i(G)$ may increase or decrease. This completes the proof to Proposition 2.3(2).

Finally, we prove Proposition 2.3(3) by construction with the following example. Consider the graph on Figure 21, and suppose that preferences are given by (2.4) with $\alpha = 1$ and $\beta = 1.5$, and assume $\gamma = 0.4$. Figure 21 shows that adding a link between a player $j \in \mathcal{S}^i(G)$ and a player $k \notin \mathcal{S}^i(G)$ can either increase (left network) or decrease (right network) $\mathcal{S}^i(G)$, and thus i's influence. ■

Figure 21: Influence and adding links



Left: effect of adding link between players 4 and 9. *Right*: effect of adding link between players 6 and 9.

Proof of Theorem 2.2

We first show that for *any* network $G(N)$, there exists a network comprising solely isolated cliques of degree k, denoted by $\mathbf{C^k}(N)$, that yields at least as great an expected contribution

level.[91] Fix γ and let i be a player with maximum expected contribution level in $G(N)$, i.e. $i \in \underset{j \in N}{\operatorname{argmax}} \left\{ \mathbb{E}_{\tilde{\theta}} \left(x_j \left(\tilde{\theta}, \gamma, G(N) \right) \right) \right\}$.

From Theorem 2.1 we know that two necessary conditions for i to contribute are that $\theta_i = \theta_R$ and that at least $\overline{c}(\gamma, k_i)$ of her neighbours are conditional cooperators, where $\overline{c}(\gamma, k_i)$ is given by (2.5). This condition is however not sufficient, as for i to contribute it is also necessary (but not sufficient) that at least $\overline{c}(\gamma, k_i)$ of her neighbours who are conditional cooperators also have $\overline{c}(\gamma, k_j)$ conditional cooperators in their own neighbourhood, and so on. Hence the following inequality:

$$\begin{aligned} \mathbb{E}_{\tilde{\theta}} \left(x_i^* \left(\tilde{\theta}, \gamma, G(N) \right) | \theta_i = \theta_R \right) &= Pr\left(i \in Q_\gamma(G) | \theta_i = \theta_R \right) \\ &\leq Pr\left(| \{ j \in N_i(G) : \theta_j = \theta_R \} | \geq \overline{c}(\gamma, k_i) \right) \end{aligned} \tag{B.5}$$

Consider next the network of isolated cliques of degree k_i, $\mathbf{C}^{\mathbf{k_i}}(N)$. Then, it follows from Theorem 2.1 that necessary and sufficient conditions for i to contribute are that $\theta_i = \theta_R$ and that at least $\overline{c}(\gamma, k_i)$ of her neighbours are conditional cooperators. The reason is that if $\theta_i = \theta_R$ and that at least $\overline{c}(\gamma, k_i)$ of i's neighbours are contributors, then these conditions also hold for her neighbours who are conditional cooperators (since i is in a clique). Hence the following equality:

$$\begin{aligned} \mathbb{E}_{\tilde{\theta}} \left(x_i^* \left(\tilde{\theta}, \gamma, \mathbf{C}^{\mathbf{k_i}}(N) \right) | \theta_i = \theta_R \right) &= Pr\left(i \in Q_\gamma \left(\mathbf{C}^{\mathbf{k_i}}(N) \right) | \theta_i = \theta_R \right) \\ &= Pr\left(| \{ j \in N_i(G) : \theta_j = \theta_R \} | \geq \overline{c}(\gamma, k_i) \right) \end{aligned} \tag{B.6}$$

Using (B.5) and (B.6), we obtain:

$$\mathbb{E}_{\tilde{\theta}} \left(x_i^* \left(\tilde{\theta}, \gamma, G(N) \right) | \theta_i = \theta_R \right) \leq \mathbb{E}_{\tilde{\theta}} \left(x_i^* \left(\tilde{\theta}, \gamma, \mathbf{C}^{\mathbf{k_i}}(N) \right) | \theta_i = \theta_R \right) \tag{B.7}$$

which completes the proof that a network of isolated cliques of degree k^* is always optimal. □

[91] For conciseness we neglect issues associated with remainder nodes when forming cliques for the proof.

We next show that $k^* \in \{\underline{k}, \underline{k}+1, \ldots n-1\}$. Recall that by definition $\underline{k}$ is the minimum degree for which there exists a regular network that supports a non-zero expected contribution level (i.e. in some ex-post ME, at least one player contributes). Hence $k^* \geq \underline{k}$, as required. ■

Proof of Proposition 2.4

To prove Proposition 2.4, we first compare the expected contribution of any player in two cliques of different sizes. We show that given social payoffs $\psi_i(.)$ that satisfy Assumption 2.1(4), either one clique is better than the other for any probability p that a player is a materialist, or there exists a unique $p^* \in (0,1)$ such that for all $p < p^*$, the larger clique is better than the smaller clique, and vice-versa for $p > p^*$. We then show that this entails that for any p, the optimal network of cliques is generically unique, and when p increases, the size of the cliques in said optimal network must (weakly) decrease.

We start with some notation. As above, denote by $\mathbb{E}_{\tilde{\theta}}\left(x_i\left(\tilde{\theta}, \gamma, \mathbf{C^k}(N)\right)\right)$ the expected contribution level per player within an isolated clique of degree k for a given γ, with:

$$\mathbb{E}_{\tilde{\theta}}\left(x_i^*\left(\tilde{\theta}, \gamma, \mathbf{C^k}(N)\right)\right) = (1-p)\, Pr\left(\left|\left\{j \in N_i\left(\mathbf{C}^k(N)\right)\theta_j = \theta_R\right\}\right| \geq \overline{c}(\gamma, k)\right) \tag{B.8}$$

where $\overline{c}(\gamma, k)$ is the minimum number of reciprocating neighbours for any conditional cooperator in the clique of degree k to cooperate.

Consider next two cliques of degrees k_A and k_B, respectively, with $k_B > k_A$. For the remainder of the proof it will be convenient to use the shorthand $\overline{c}(\gamma, k_A) \equiv c_A$ and $\overline{c}(\gamma, k_B) \equiv c_B$. Assumption 2.1(4) implies that $c_B \geq c_A$. If $c_A = c_B$, then the larger clique clearly has higher expected contributions than the smaller one. Another implication of Assumption 2.1(4) is that $k_B - c_B \geq k_A - c_A$. If equality holds, i.e. $k_B - c_B = k_A - c_A$, then the larger clique can tolerate no more materialists than the smaller one consistent with its players cooperating. As such the smaller clique trivially dominates the larger one for any p. It thus remains to

consider the case that $c_A < c_B$ and $k_B - c_B > k_A - c_A$.

Let Y_k be the number of conditional cooperators from k draws, so that $Y_{k,1-p} \sim \text{Bin}(k; 1-p)$. Let $F_{k,1-p}$ denote the cumulative distribution of $Y_{k,1-p}$, so that

$$\mathbb{E}_{\tilde{\theta}}\left(x_i^*\left(\tilde{\theta}, \gamma, \mathbf{C^k}\left(N\right)\right)\right) = (1-p)\left(1 - F_{k-1,1-p}(\overline{c}(\gamma,k)-1)\right) \tag{B.9}$$

For convenience we write the cumulative distribution $F_{k-1,1-p}(\overline{c}(\gamma,k)-1)$ in terms of a beta function, as follows:

$$F_{k-1,1-p}(\overline{c}(\gamma,k)-1) = (k-\overline{c}(\gamma,k))\begin{pmatrix} k-1 \\ \overline{c}(\gamma,k)-1 \end{pmatrix}\int_{u=0}^{p} u^{k-\overline{c}(\gamma,k)-1}(1-u)^{\overline{c}(\gamma,k)-1}du \tag{B.10}$$

Next, consider the following ratio:

$$R\left(p\right) \equiv \frac{F_{k_B-1,1-p}(c_B-1)}{F_{k_A-1,1-p}(c_A-1)} = C\frac{I_B(p)}{I_A(p)} \tag{B.11}$$

where C is a constant with:

$$C \equiv \frac{(k_B-\overline{c}(\gamma,k_B))\begin{pmatrix} k_B-1 \\ \overline{c}(\gamma,k_B)-1 \end{pmatrix}}{(k_A-\overline{c}(\gamma,k_A))\begin{pmatrix} k_A-1 \\ \overline{c}(\gamma,k_A)-1 \end{pmatrix}} > 1 \tag{B.12}$$

and:

$$I_B(p) = \int_{u=0}^{p} u^{k_B-c_B-1}(1-u)^{c_B-1}du \tag{B.13}$$

and $I_A(p)$ defined analogously. Note that this ratio can be used to determine which of the cliques A and B yields a higher ex-ante contribution level per player. In particular, when $R\left(p\right) < 1$, the (larger) clique of degree k_B yields higher ex-ante contribution level per player than the (smaller) clique of degree k_A; the converse holds true for any $R\left(p\right) > 1$.

Lemma B.1 *For any two cliques of degree k_A and k_B, with $k_B > k_A$ and $k_B - c_B > k_A - c_A$, there exists a unique $p^* \in (0,1)$ such that for $R(p)$ given by (B.11), $R(p^*) = 1$, with $R(p^*) < 1$ for all $p < p^*$ and $R(p^*) > 1$ for all $p > p^*$.*

Proof: The proof has the following structure. We establish three properties of $R(p)$:

1. $R(p) \to 0$ as $p \to 0$;
2. $R(1) = 1$;
3. $R(p)$ is strictly increasing up to some unique $\hat{p} \in (0,1)$ and is thereafter strictly decreasing.

Note that Properties 2 and 3 together imply that $R(p)$ increases up to a point $\hat{p}$, where it attains its unique maximum $R(\hat{p}) > 1$, and further that $R(p) > 1$ for all $p \in [p', 1)$. Property 1 then implies the lemma straightforwardly through a fixed-point argument.

We first prove the first property. Clearly, as $p \to 0$, $I_B(p) \to 0$ and $I_A(p) \to 0$. Also, by definition we know that both $I_B(p)$ and $I_A(p)$ are continuous and strictly increasing on the domain $(0,1]$. By the fundamental theorem of calculus, we have that

$$\frac{\partial [I_A(p)]}{\partial p} = p^{k_A - c_A - 1}(1-p)^{c_A - 1} \tag{B.14}$$

$$\frac{\partial [I_B(p)]}{\partial p} = p^{m_1}(1-p)^{m_2} p^{k_A - c_A - 1}(1-p)^{c_A - 1} \tag{B.15}$$

where m_1 and m_2 are strictly larger than 1. Thus, L'Hôpital's rule establishes the following:

$$\begin{aligned} \lim_{p \to 0} \frac{I_B(p)}{I_A(p)} &= \lim_{p \to 0} \frac{p^{m_1}(1-p)^{m_2} p^{k_A - c_A - 1}(1-p)^{c_A - 1}}{p^{k_A - c_A - 1}(1-p)^{c_A - 1}} \\ &= \lim_{p \to 0} p^{m_1}(1-p)^{m_2} \\ &= 0 \end{aligned} \tag{B.16}$$

which proves the first property. Property 2 is straightforwardly established by noting that when $p = 1$, $F_{k_A - 1, 1-p}(c_A - 1) = 1$ and $F_{k_B - 1, 1-p}(c_B - 1) = 1$. Hence, when $p = 1$, $R(p) = 1$.

We now turn to Property 3. We first show that $\frac{I_B(p)}{I_A(p)}$ and thus $R\left(p\right)$ are strictly increasing on $(0,\hat{p})$, for some $\hat{p}\in(0,1)$. Consider the following derivative:

$$\frac{\partial}{\partial p}\left(\frac{I_B(p)}{I_A(p)}\right) = \frac{I_B(p)}{I_A(p)}\left(\frac{p^{k_B-c_B-1}(1-p)^{c_B-1}}{I_B(p)}-\frac{p^{k_A-c_A-1}(1-p)^{c_A-1}}{I_A(p)}\right) \tag{B.17}$$

which we can express as

$$\begin{gathered}\frac{\partial}{\partial p}\left(\frac{I_B(p)}{I_A(p)}\right)= \\ \frac{I_B(p)}{I_A(p)}\left(\frac{(p^{m_1}(1-p)^{m_2})p^{k_A-c_A-1}(1-p)^{c_A-1}}{\int_{u=0}^{p}(u^{m_1}(1-u)^{m_2})u^{k_A-c_A-1}(1-u)^{c_A-1}du}-\frac{p^{k_A-c_A-1}(1-p)^{c_A-1}}{\int_{u=0}^{p}u^{k_A-c_A-1}(1-u)^{c_A-1}du}\right)\end{gathered} \tag{B.18}$$

and so $\frac{d}{dp}\left(\frac{I_B(p)}{I_A(p)}\right)>0$ if and only if $g(p)>0$,where:

$$g(p):=m\left(p\right)\int_{u=0}^{p}u^{k_A-c_A-1}(1-u)^{c_A-1}du-\int_{u=0}^{p}m(u)u^{k_A-c_A-1}(1-u)^{c_A-1}du \tag{B.19}$$

where, in turn,

$$m\left(p\right)=p^{m_1}(1-p)^{m_2} \tag{B.20}$$

Note that the function $m\left(p\right)$ is strictly increasing up to a maximum at $p'=\frac{m_1}{m_1+m_2}$. For $p\leq\frac{m_1}{m_1+m_2}$, then, it follows that $m\left(u\right)<m\left(p\right)$ for all $u<p$, which ensures that $g(p)>0$. Hence, $\frac{I_B(p)}{I_A(p)}$ is strictly increasing on the interval $(0,p']$.

Next, consider the interval $(p',1]$. On this interval we have that

$$\begin{aligned}\frac{\partial g(p)}{\partial p} &= \frac{\partial m(p)}{\partial p}\int_{u=0}^{p} u^{k_A-c_A-1}(1-u)^{c_A-1}du + m\left(p\right)p^{k_A-c_A-1}(1-p)^{c_A-1} \\ &\quad - m\left(p\right)p^{k_A-c_A-1}(1-p)^{c_A-1} \\ &= \frac{\partial m(p)}{\partial p}\int_{u=0}^{p} u^{k_A-c_A-1}(1-u)^{c_A-1}du \\ &< 0\end{aligned} \tag{B.21}$$

where the inequality follows from the fact that $m\left(p\right)$ is strictly decreasing on this interval. Further, we know that $g(p) < 0$ for a high enough probability p, since the term $\int_{u=0}^{p} m(u)u^{k_A-c_A-1}(1-u)^{c_A-1}du$ increases continuously to a positive finite number as $p \to 1$ while the term $m\left(p\right)\int_{u=0}^{p} u^{k_A-c_A-1}(1-u)^{c_A-1}du$ tends to zero. Hence we know that there must exist a unique $\hat{p} \in (p', 1)$ such that $g(\hat{p}) = 0$. It follows that $\frac{I_B(p)}{I_A(p)}$ (and hence $R\left(p\right)$) is strictly increasing on $(0, \hat{p}]$ and strictly decreasing on $(\hat{p}, 1)$. This completes the proof of the third property.

It follows from the three properties above that $R\left(p\right) \to 0$ as $p \to 0$, that $R\left(p\right)$ strictly increases on the interval $(0, \hat{p}]$, reaches a maximum at $p = \hat{p}$ and then is strictly decreasing on $[\hat{p}, 1)$, attaining the value of unity at $p = 1$. It follows from a standard fixed point argument that there must exists a unique $p^* \in (0, 1)$, such that $R\left(p^*\right) = 1$, $R\left(p\right) > 1$ for $p > p^*$ and $R\left(p\right) < 1$ for $p < p^*$. This completes the proof to Lemma B.1. □

We have shown that for any two pair of cliques of degree k_A and k_B, with $k_B > k_A$, either one clique is better than the other for any probability p that a player is a materialist, or there exists a unique $p^* \in (0, 1)$ such that for all $p < p^*$, the larger clique is better than the smaller clique, and vice-versa for $p > p^*$. Proposition 2.4 follows immediately. ■

B.2 Extension: repeated interactions

In this section, we show that our model can be seen as the reduced form of an infinitely repeated local public good game where players have only material payoffs. We demonstrate that for any ME $\mathbf{x}^*$ of the one-shot model, there exists a closely related stage game among materialist players such that $\mathbf{x}^*$ is the maximal equilibrium that can be sustained in the infinitely repeated version of the stage game. We also relax Assumption 2.1 in an important way by allowing free-riding players to have social payoffs.

Game and strategies. Let $N = \{1, 2, ...n\}$ be the set of players, with $n \geq 3$. Denote by G an (undirected and unweighted) network, as before. A player i's degree is denoted $k_i \in \left[0, \overline{k}\right]$, where $\overline{k}$ is the maximal degree in N. We assume that the network does not change over time. Agents play an infinitely-repeated local public good game in discrete time. In period t, player i chooses whether to cooperate ($x_{it} = 1$) or not ($x_{it} = 0$). $X_i^t = \{0, 1\}$. For a given profile $\mathbf{x_t}$, c_{it} and d_{it} denote respectively the number of i's neighbours who contribute and free-ride in period t.

Denote by $\mathbf{h}_t$ the history of play at the end of period t (i.e. after players have played their action for that period), where $\mathbf{h}_t = (\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_{t-1}})$ At period t, players thus condition their play on $\mathbf{h}_t$: we denote a player i's strategy by $\sigma_i(\mathbf{h})$, where $\mathbf{h}$ is a history of arbitrary length. Let $\tilde{\mathbf{x}}_t(\sigma)$ be the *outcome at* t, i.e. the action profile induced at t by strategy profile σ.

Payoffs. At a given period t, a player i's stage payoffs for a given profile $\mathbf{x}_t$ are given by:

$$v\left(x_{it}, \mathbf{x_{-it}}\right) = \varphi\left(x_{it}, c_{it}, d_{it}\right) - \gamma x_{it} \tag{B.22}$$

where $\gamma > 0$ is the net cost of contribution, as before; and where $\varphi\left(\cdot\right)$ represents players' payoffs to the local public good. Define:

$$\phi\left(c_{it}, d_{it}\right) = \varphi\left(1, c_{it}, d_{it}\right) - \varphi\left(0, c_{it}, d_{it}\right) \tag{B.23}$$

Assumption B.1 *Players' stage-game payoffs are given by (B.22), where:* ***(1)*** $\varphi\left(x_i, 0, k_i\right) = 0$ *for any* $k_i \in \left[0, \overline{k}\right]$*;* ***(2)*** $\varphi\left(1, k_i, 0\right) - \gamma < \varphi\left(0, k_i, 0\right)$ *for any* $k_i \in \left[0, \overline{k}\right]$*;* ***(3)*** $\varphi\left(x_{it}, c_{it}, d_{it}\right)$ *is*

weakly increasing (decreasing) in c_{it} (d_{it}); and ***(4)*** *$\phi(c_{it}, d_{it})$ is strictly increasing (decreasing) in c_i (d_i).*

We briefly comment on Assumption B.1. Assumption B.1(1) is a normalisation. Assumption B.1(2) guarantees that free-riding is always optimal in the stage-game. Assumption B.1(3) captures the idea of a local public good: a player i incurs a positive (negative) externality on her neighbours when she cooperates (free-rides). Lastly, Assumption B.1(4) states that cooperation exhibits *strategic complementarity*. This assumption may capture different intuitions, e.g. the cost of cooperating decreases with local cooperation because of learning opportunities; neighbours' cooperation or effort increases the marginal returns to own cooperation or effort; etc. Observe that Assumptions B.1(1), B.1(3) and B.1(4), together, form a generalised version of Assumption 2.1.

We write player i's continuation payoffs at period t as:

$$u_{i,t}\left(\{\mathbf{x_{ik}}\}_{k=t}^{\infty}\right) = \sum_{k=t}^{\infty} \left[\delta\left(\theta_i\right)\right]^k v\left(x_{it}, \mathbf{x_{-it}}\right) \tag{B.24}$$

where $\delta(\theta_i)$ is player i's *discount factor*, which depends on her type. Player i's type $\theta_i \in \Theta_i = \{\theta_F, \theta_M\}$ is ascribed by nature at the beginning of the game, with θ_F and θ_M referring respectively to *forward-looking* and *myopic* types. We assume that $\delta(\theta_F) = \delta$, with $\delta \in (0,1)$, and $\delta(\theta_M) = 0$. Hence, while forward-looking players play "the long game", myopic players play the infinite game at each period as if it were only the stage-game. These types form a close analogy to the "conditional cooperator" and "materialist" types introduced earlier.

Equilibrium. A strategy profile σ^* is a (subgame-perfect) equilibrium if at no period $t \geq 1$ does there exist a strategy σ' such that, for some player i, $u_i^t\left(\sigma', \sigma_{-i}^*\right) > u_i^t\left(\sigma_i^*, \sigma_{-i}^*\right)$. An equilibrium σ^* is *maximal* if, at any $t \geq 1$ and for any other strategy σ':

$$\sum_{i \in N} \tilde{x}_i^t(\sigma^*) \geq \sum_{i \in N} \tilde{x}_i^t(\sigma') \tag{B.25}$$

It is straightforward to see that at any equilibrium, all myopic players defect at every period. However, forward-looking players may achieve and enforce some cooperation at equilibrium, depending on their strategy. For simplicity, we restrict attention to equilibria supported by *local trigger strategies*.

Assumption B.2 *All forward-looking players adopt a* local trigger strategy*, whereby any forward-looking player i's strategy $\sigma_i(\mathbf{h})$ is such that: $x_{i,t} = 1$ if for every $k \in \{1, t-1\}$, $x_{j,k} = 1$ for at least μ_i neighbours $j \in N_i(G)$; and $x_{i,t} = 0$ otherwise.*

Assumption B.2 states that any forward-looking player i cooperates as long as in every previous period, at least μ_i neighbours cooperated. We do not specify μ_i, and so Assumption B.2 remains fairly permissive.

We now examine the maximal equilibrium that can be supported when forward-looking players adopt a grim trigger strategy as above. Before stating our result, we adapt our definition of the q-linked set to the repeated environment. For any $q \in \mathbb{R}_+$, we define the *generalised q-linked set* of G as the largest set of players such that for each i in the repeated q-linked set:

$$\phi\left(s_i^q, (k_i - s_i^q)\right) + \delta\varphi\left(0, s_i^q, (k_i - s_i^q)\right) \geq q \tag{B.26}$$

where s_i^q is, as before, the number of i's neighbours in the repeated q-linked set. Observe that when $\delta = 0$, condition (B.26) suitably mirrors condition (2.6) from the benchmark model.

Theorem B.1 *Suppose that Assumptions B.1 and B.2 hold. For any $\gamma \in \mathbb{R}_+$, G and θ, a ME always exists and is unique. At the ME, a forward-looking player contributes at every period if and only if she is in the generalised γ-linked set. A local trigger strategy that supports the ME is such that for every forward-looking player i, $\mu_i = s_i^q$.*

Proof: The proof of Theorem B.1 follows closely the proof of Theorem 2.1. Suppose that all players in the generalised γ-linked set cooperated in every period $k \in \{1, 2...t-1\}$: observe

that when $\mu_i = s_i^q$, then a forward-looking player i in the generalised q-linked set finds it profitable to continue to cooperate at time t if:

$$\frac{\varphi(1, c_i, d_i) - \gamma}{1 - \delta} \geq \varphi(0, c_i, d_i) + \frac{\delta\varphi(0, 0, k_i)}{1 - \delta} \tag{B.27}$$

The LHS of condition (B.27) captures i's discounted future payoffs from cooperating, while the RHS captures i's payoffs from deviating (which entails that all i's neighbours defect forever from the next period onwards). It is easy to show using condition (B.26) that if i is in the generalised q-linked set, then condition (B.26) is satisfied. Hence, when $\mu_i = s_i^q$, all players in (outside) the generalised q-linked set find that cooperating (free-riding) is optimal.

We next sketch the proof to the claim that this is indeed the *maximal* equilibrium. Suppose *a contrario* that there exists another equilibrium in local trigger strategies σ' such that, at some $t \geq 1$, $\sum_{i \in N} \tilde{x}_{it}(\sigma') > \sum_{i \in N} \tilde{x}_{it}(\sigma^*)$. Then, the set of players C_t contributing at t on the equilibrium path of σ' must be strictly larger than the generalised q-linked set. In that case, for each player $i \in C_t$,

$$\phi(\hat{s}_i^q, (k_i - \hat{s}_i^q)) + \delta\varphi(0, \hat{s}_i^q, (k_i - \hat{s}_i^q)) \geq \gamma \tag{B.28}$$

where $\hat{s}_i^q$ is the proportion of i's neighbours that are members of C_t. But, by construction, the generalised q-linked set is the unique largest set for which this is true. This contradicts the claim that C_t is larger than the generalised q-linked set. ■

We conclude this section with a few remarks. First, Theorem B.1 establishes that the results of our benchmark game can be seen as a reduced form of a repeated game where players have only material payoffs but display heterogeneous discount factors. Theorem B.1 also shows that our results are robust to the generalisation of Assumption 2.1 embedded in Assumption B.1. Lastly, in the repeated setting, "cooperation" is never optimal in the stage-game (as opposed to our benchmark game). This extension thus yields a different interpretation when interactions are repeated: cooperation can be sustained by players in the generalised q-linked set because it is enforced by threat of punishment, implicit in trigger strategies.

B.3 Extension: private types

In our benchmark model, we assume that players know the type profile. Here we assume that players do not know each other's type, and types are i.i.d. over (θ_M, θ_R) with probability distribution $(p, 1-p)$. We first construct a particular BNE that can be characterised via a set $\hat{\mathcal{Q}}_\gamma$, which is analogous to the γ-linked set $\mathcal{Q}_\gamma$. We then show that this BNE must be the unique maximal BNE.

In this setting, a (pure) strategy for player i, denoted $y_i(.)$, is a mapping $\Theta_i \to \{0,1\}$ from i's type to i's action. Denote by Y_i the set of all (four) possible such mappings. Fix a set of players N and a network G. A strategy profile for all n players is denoted $\mathbf{y}$; a type profile is denoted $\theta \in \Theta = \{\theta_M, \theta_R\}^n$. Denote by $\theta_{j\in N_i(G)}$ the *local type profile*, i.e.. the type profile among player i's neighbours. We first make the two following definitions.

Definition B.1 *A Bayesian Nash Equilibrium (BNE) is a strategy profile* $\mathbf{y}$ *for all* n *players such that, for every* $i \in N$ *and every* $y_i' \in Y_i$, $E_{\theta_{-i}}\left[\pi\left(y_i, \mathbf{y}_{-i}|\theta_i\right)\right] \geq E_{\theta_{-i}}\left[\pi\left(y_i', \mathbf{y}_{-i}|\theta_i\right)\right]$.

Definition B.2 *A BNE* $\mathbf{y}^*$ *is* maximal *if there does not exist another BNE* $\mathbf{y}' \in \{0,1\}^n$ *such that* $E_\theta\left[\sum_{i\in N} y_i^*(\theta_i)\right] < E_\theta\left[\sum_{i\in N} y_i'(\theta_i)\right]$.

Next, we define the expected net benefit from contributing if $\theta_i = \theta_R$ as follows.

$$E_{\theta_{-i}}\left[\Delta_i\left(\mathbf{y}_{\mathbf{j}\in\mathbf{N_i(G)}}|\theta_i\right)\right] = \sum_{\theta\in\Theta} f(\theta)\Delta_i\left(\mathbf{y}_{j\in N_i(G)}\left(\theta_{j\in N_i(G)}\right)|\theta_i\right) \tag{B.29}$$

where $f(\cdot)$ is the probability distribution over local type profiles, and $\Delta_i\left(\mathbf{y}_{j\in N_i(G)}|\theta_i\right)$ is given by:

$$\Delta_i\left(\mathbf{y}_{j\in N_i(G)}|\theta_i\right) = -\gamma + \Psi\left(x_i, c_i, d_i|\theta_i\right)$$

Finally, for any $q \in \mathbb{R}_+$, we define the *Bayesian q-linked set*, $\hat{\mathcal{Q}}_q$, to be the largest set of players for whom:

$$E_{\theta_{-i}}[\Delta_i\left(\mathbf{y}_{\mathbf{j}\in\mathbf{N_i(G)}^*}|\theta_i\right)] \geq q \tag{B.30}$$

for each player $i \in \hat{\mathcal{Q}}_q$, where the profile $\mathbf{y}^*$ is defined to be $(y_j^*(\theta_M), y_j^*(\theta_R)) = (0,1)$ if $j \in \hat{\mathcal{Q}}_\gamma$ and $(y_j^*(\theta_M), y_j^*(\theta_R)) = (0,0)$ otherwise.

Theorem B.2 *Suppose that Assumption 2.1 holds and that players' types are private and i.i.d., with* $p \equiv Pr(\theta_i = \theta_M) \in (0,1)$ *for all* $i \in N$. *For any* $\gamma \in \mathbb{R}_+$, G *and* p, *a maximal BNE always exists and is unique. At the maximal BNE, a player contributes if and only if she is in the Bayesian* γ*-linked set.*

Proof: Analogously to the proof of Theorem 2.1. The profile $\mathbf{y}^*$ is clearly a BNE for any $\gamma \in \mathbb{R}$. It also follows from the proof to Theorem 2.1 that this maximal BNE is unique, which is implied by the fact that $\hat{\mathcal{Q}}_\gamma$ is the unique maximal set containing only members for whom inequality B.30 holds. ■

Remark B.1 *Membership of the Bayesian* γ*-linked set entails that a player contributes only if she is a conditional cooperator type; players not in the set always free-ride. The expected contribution level at the maximal BNE is thus* $(1-p) \mid \hat{\mathcal{Q}}_\gamma \mid$.

The close similarity between the definition of $\hat{\mathcal{Q}}_q$ and that of $\mathcal{Q}_q$ ensures that the comparative statics result from our main model continue to hold in the case of private types. First, consider an increase in γ. This makes membership of $\hat{\mathcal{Q}}_\gamma$ more stringent, thus reducing expected contributions. Second, consider adding a link to a network between a player $i \in \hat{\mathcal{Q}}_\gamma$ and a player $j \notin \hat{\mathcal{Q}}_\gamma$. We now argue that, just as in the case of public types, the effect of adding this link can increase or decrease expected contributions (or keep it the same). If i only narrowly meets inequality B.30, while j is far from meeting it, then the result will be that i switches to free-riding regardless of type. In this case, expected contributions decrease. On the other hand, if i comfortably meets inequality B.30, while j just fails to meet it initially, then adding a link increases expected contributions. Finally, if neither player is 'marginal' with respect to inequality (B.30), then adding the link does not affect expected contributions.

C Appendix to Chapter 3

C.1 Proofs

Proof of Proposition 3.1

I first prove that FH implies that strong ideology does not hold. Fix arbitrary values of $\mathbf{z} \in Z^n$, $\theta, \theta' \in \Theta$, where $\theta \neq \theta'$. Consider $\tilde{\theta}_{-i}, \hat{\theta}_{-i} \in \Theta^{n-1}$ such that $\mathcal{C}_\theta(\tilde{\theta}_{-i}) > \mathcal{C}_\theta(\hat{\theta}_{-i})$ and $\mathcal{C}_{\theta'}(\tilde{\theta}_{-i}) < \mathcal{C}_{\theta'}(\hat{\theta}_{-i})$. FH implies that $w_\theta(\mathbf{z}; \theta', \theta_{-i})$ is increasing in $\mathcal{C}_{\theta'}(\theta', \theta_{-i})$, which implies that $w_\theta\left(\mathbf{z}; \theta', \tilde{\theta}_{-i}\right) < w_\theta\left(\mathbf{z}; \theta', \hat{\theta}_{-i}\right)$. SI implies $\min_{\mathbf{z} \in Z^n} w_\theta\left(\mathbf{z}; \theta', \tilde{\theta}_{-i}\right) > \max_{\mathbf{z} \in Z^n} w_\theta\left(\mathbf{z}; \theta', \hat{\theta}_{-i}\right)$, which in turn implies $w_\theta\left(\mathbf{z}; \theta', \tilde{\theta}_{-i}\right) > w_\theta\left(\mathbf{z}; \theta', \hat{\theta}_{-i}\right)$.

It remains to prove that strong and weak ideology are consistent with SI. Continuing to fix arbitrary parameter values as above, SI implies that $\min_{\mathbf{z} \in Z^n} w_\theta\left(\mathbf{z}; \theta, \tilde{\theta}_{-i}\right) > \max_{\mathbf{z} \in Z^n} w_\theta\left(\mathbf{z}; \theta, \hat{\theta}_{-i}\right)$, which in turn implies $w_\theta\left(\mathbf{z}; \theta, \tilde{\theta}_{-i}\right) > w_\theta\left(\mathbf{z}; \theta, \hat{\theta}_{-i}\right)$. Since $\tilde{\theta}_{-i}, \hat{\theta}_{-i} \in \Theta^{n-1}$ are arbitrarily fixed given the constraint that $\mathcal{C}_\theta(\tilde{\theta}_{-i}) > \mathcal{C}_\theta(\hat{\theta}_{-i})$, this implies that $w_\theta(\mathbf{z}; \theta, \theta_{-i})$ is increasing in $\mathcal{C}_\theta(\theta_{-i})$, i.e. PH. Hence strong ideology implies PH. Weak ideology of type θ implies that there exists some $\theta' \in \Theta$ such that $w_\theta(\mathbf{z}; \theta', \theta_{-i})$ is not strictly increasing in $\mathcal{C}_{\theta'}(\theta_{-i})$. This does not however preclude that $w_\theta(\mathbf{z}; \theta, \theta_{-i})$ is increasing in $\mathcal{C}_\theta(\theta_{-i})$, i.e. that PH holds. For example, a player of weakly ideological type A facing a type-B opponent could have $w_A(\mathbf{z}; B, A) = w_A(\mathbf{z}; B, B)$ but $w_A(\mathbf{z}; A, A) > w_A(\mathbf{z}; A, B)$. ■

Proof of Theorem 3.1

First suppose A and B are ideological. For arbitrarily small $c > 0$, $\Delta^{retain}_{A,B} > c$ and $\Delta^{retain}_{B,A} > c$, and so there exists an equilibrium $\mathbf{x}^* = (1, 1)$ such that $E(\mathbf{x}^*) = -2c < 0$. This proves Theorem 3.1(1). Turning to the next part of the Theorem, if both types are strongly ideological then for arbitrarily small $c > 0$, $\Delta^{convert}_A > c$, $\Delta^{convert}_B > c$, $\Delta^{retain}_{A,B} > c$ and $\Delta^{retain}_{B,A} > c$, implying $\mathbf{x}^* = (1, 1)$ and $E(\mathbf{x}^*) = -2c < 0$. Finally, to prove Theorem 3.1(3) let A be perfectly ideological and let B be pragmatic. For arbitrarily small $c > 0$, $\Delta^{convert}_A > c$. As $\Delta^{retain}_{B,A} < 0 < c$, $\mathbf{x}^* = (1, 0)$ and $E(\mathbf{x}^*) > 0$. ■

C.2 Characterisation of equilibria in case 3 of extended example in section 3.2

Recall that having weakly ideological meta-payoffs entails that a player always prefers to retain her type, regardless of any change in outcome in Γ that a change in her type would induce. Additionally, for the particular weakly ideological types in the example, no players derive utility directly from their opponents' types. Turning to equilibrium analysis, we see that if no devout players invest, then secular players all best-respond if none invests. In contrast, however, if no secular players invest, then provided investment costs are low enough ($c < \frac{n_s}{n}$) devout players each best-respond if precisely one of them invests. This response is driven by the instrumental incentive to convert their opponents to improve the outcome they achieve in Γ. Now suppose that $k > 0$ devout players invest (where $k \leq n_d$). In this case, secular players each best-respond if k of them invest, provided $c < 2$ (as their meta-payoff if they retain their type is 2 higher than that if they are converted), which must be true if $c < \frac{n_s}{n}$. If k secular players invest (where $k \leq n_s$), then devout players each best-respond if (i) $k+1$ of them invest when $c < \frac{n_s}{n}$; (ii) k of them invest when $c \in (\frac{n_s}{n}, 2)$ and (iii) none of them invest when $c > 2$. Note that an investor among a total of $k+1$ devout investors does not deviate if the cost of investing is less than the instrumental incentive to convert (case (i)), whereas the ideological incentive is relevant to an investor among a total of k devout investors (case (ii)), because deviating here will ensure that the devout investor changes type. It follows that:

1. If $c > 2$, no players invest in equilibrium
2. If $c \in (\frac{n_s}{n}, 2)$, an equilibrium investment profile is any in which precisely m devout and m secular players invest, where $m \in \{0, 1, \ldots, \min\{n_s, n_d\}\}$
3. If $c < \frac{n_s}{n}$ and $n_d \leq n_s$, then an equilibrium investment profile is any in which precisely n_d devout and n_d secular players invest
4. If $c < \frac{n_s}{n}$ and $n_d > n_s$, then there is no pure-strategy equilibrium investment profile

The extended motivating example of the ascetic religion, and the three different cases considered above help to illustrate how meta-payoffs can be used to characterise different kinds of ideology, and how these kinds of ideology affect strategic incentives. Insights from the main results are summarised as follows. In the first case, where a weakly ideological devout type meets a pragmatic secular type, there is a threshold value for the investment cost c below which precisely one player, of the ideological type, invests in equilibrium. Turning to the second case, if the devout type is instead strongly ideological, this simply strengthens the incentive of devout players to convert, as their instrumental incentive to convert is coupled with an ideological incentive to convert, and so the result is much the same as in the first case but with a reduced threshold cost.

C.3 Example application: selection of meta-preferences

In this section of Appendix C, I consider how the indirect evolutionary model of Chapter 1 can be extended to ideological games. In applying ideological games to this context, I provide an account specifically of cultural – as opposed to genetic – evolution.

There is a continuum of players, indexed by the interval $[0, 1]$ containing two preference types: materialists (M), whose (first-order) payoffs are simply the fitness payoffs, and reciprocators (R), who rank mutual cooperation above the other outcomes of the game (see Figures 23 and 24). Players undergo uniform random pairing to play a game Γ that is a PD in fitness payoffs, as given in Figure 22. Assuming players have complete information (i.e. preference types are observable), reciprocators can thrive. A further possibility – and one that has not been formally studied, to my knowledge – is that selection can take place over meta-preferences. Ideological games offer a means to study such a mechanism. Specifically, they enable the study of equilibrium behaviour as resulting from 'competition' between ideologies. In this example, Γ is the second and final stage of a two-player ideological game. The example shows that, just as first-order preference types can be explained via the indirect evolutionary approach, using the framework studied in Chapter 1 of this thesis, meta-preference

types – and specifically, ideologies – can be explained by extending the indirect evolutionary approach to ideological games. In particular, the example sets out how ideology can allow materialist players to gain higher fitness than non-ideological reciprocators at some population shares, whereas in the absence of such ideology, materialist players suffer lower fitness at all population shares, assuming complete information. If reciprocators also have ideological meta-preferences, however, the range of population shares at which materialists enjoy a fitness advantage can be decreased.

In any given pair, if both players invest or neither invests, then the final type profile is unchanged from the initial type profile. If one player invests but the other does not, then in the final type profile the non-investing player is converted to the type of the investing player. Formally, the conversion technology is that given by(3.21) in the two-player setting, i.e.

$$f(\theta_1, \theta_2, x_1, x_2) = \begin{cases} (\theta_1, \theta_2) & if\ x_1 = x_2 \in \{0, 1\} \\ (\theta_1, \theta_1) & if\ (x_1, x_2) = (1, 0) \\ (\theta_2, \theta_2) & if\ (x_1, x_2) = (0, 1) \end{cases} \tag{C.1}$$

One difference between the present context and that in Chapter 1 is that the effect of conversions on players' fitness needs to be taken into account. A simple way to do this is to assign a negative fitness payoff to the converted player, alongside the resulting fitness payoff from Γ. An interpretation of this approach is that while a player can convert their opponent before the two players play Γ, in the future the opponent may subsequently revert back to their original type. While the probability with which such a reversion takes place is not specified, the effect is picked up in the fitness by the negative payoff associated with being converted. Another complication is that, if conversion takes place, the fitness payoff from Γ the converted player receives is in fact a fitness payoff for the player's new type. However, we can think of this as also being taken into account by the negative fitness payoff assigned to the converted player. I will assign a value of -1 to the fitness of a converted player in this example.

The subgame of Γ in which two reciprocators meet (denoted Γ_{RR}) forms a coordination

Figure 22: Γ forms a PD in fitness payoffs

	$z_j = 1$	$z_j = 0$
$z_i = 1$	(3,3)	(-1,4)
$z_i = 0$	(4,-1)	(0,0)

Figure 23: Γ forms a coordination game in subjective payoffs when two type-R players meet

	$z_j = 1$	$z_j = 0$
$z_i = 1$	(5,5)	(-1,4)
$z_i = 0$	(4,-1)	(0,0)

game in subjective payoffs. Assumption 3.2 requires that we specify which of the two Nash equilibria is played; let us suppose it is $\mathbf{z}^* = (1, 1)$, inducing the top-left outcome in the table. Finally, Γ_{MR} is the game of Figure 24 (with Γ_{RM} symmetric), which clearly has the unique Nash equilibrium $\mathbf{z}^* = (0, 0)$.

In other words, the type-R (reciprocator) player gains utility in line with fitness payoffs except for in the case of mutual cooperation, in which case she has a subjective payoff exceeding all her other possible subjective payoffs in Γ. Suppose that types are common knowledge, and there is uniform random matching into pairs, among a large population of players, to play Γ. In such a setting, in the absence of second-order preferences reciprocators do better than materialists, since they can cooperate when they meet one another but defect when they encounter materialists.

Benchmark result (Dekel, Ely and Yilankaya, 2006): *In a population containing types M and R only, R gains higher expected fitness than type M.*

This difference in fitness will depend on the population shares of the types, denoted ε_M and ε_R for M and R respectively, so that $\varepsilon_M + \varepsilon_R = 1$. Recall that the fitness payoffs associated with Γ are those of the PD in Figure 22. The dominant strategy in Γ for a type-M player is to defect, i.e. play $z = 0$. For type-R players, by assumption $\mathbf{z}^* = (1, 1)$ is played in Γ_{RR}, while on meeting a materialist, the best response for a type-R player is to play $z = 0$. Consequently, type M players receive zero expected fitness, since any type-M player defects against a defecting opponent. For type-R players, expected fitness is $3\varepsilon_R \geq 0$, since

Figure 24: Subjective payoffs when type M (row player) meets type R (column player)

	$z_j = 1$	$z_j = 0$
$z_i = 1$	(3,5)	(-1,4)
$z_i = 0$	(4,-1)	(0,0)

their fitness payoff from Γ is zero if paired with a materialist and 3 if paired with a fellow reciprocator, which happens with probability $\varepsilon_R \geq 0$.

Continuing to develop the application, let us now suppose that in the initial stage of the two-player ideological game, investment now generates a fitness cost of c, equalling the disutility it entails for either type of player. For an ideological game to be well-defined, it is necessary to specify meta-payoffs for each type. Suppose meta-payoffs now have the following properties.[92]

- M' is (strongly) ideological, with $\Delta_{M'}^{conv\,(id)} = 6$, $\Delta_{M'}^{conv\,(in)} = 0$, $\Delta_{M',R}^{ret\,(id)} = 6$, $\Delta_{M',R}^{ret\,(in)} = 0$
- R is pragmatic, with $\Delta_{R}^{conv\,(id)} = 0$, $\Delta_{R}^{conv\,(in)} = 5$, $\Delta_{R,M'}^{ret\,(id)} = 0$, $\Delta_{R,M'}^{ret\,(in)} = 0$

Assumption 3.7 ensures meta-payoffs replicate first-order payoffs in the event players retain their type. This ensures that the instrumental incentives to convert and retain follow from the specifications of first-order preferences illustrated in Figures 22 to 24. In contrast, the ideological incentives are specified here for the first time.

In the case of the instrumental incentives to convert, for a materialist, equilibrium payoffs in Γ are the same regardless of the opponent's type, so $\Delta_{M'}^{conv\,(in)} = 0$. Reciprocators, on the other hand, achieve a better Nash outcome in Γ when facing a same-type opponent, since $u_R(\mathbf{z}^*(R,R)) = u_R(1,1) = 5 > 0 = u_R(0,0) = u_R(\mathbf{z}^*(R,M))$.[93] By Lemma 3.1, we can write $w_R(\mathbf{z};\theta,\theta_{-i}) = u_R(\mathbf{z}) + v_R(\theta_{-i})$. Hence $\Delta_{R'}^{conv\,(in)} = u_R(\mathbf{z}^*(R,R)) - u_R(\mathbf{z}^*(R,M)) = 5$. Let us suppose that M' has the first-order payoffs of type M.[94] In this case, the instrumental incen-

[92]The notation follows that in (3.36) and (3.37).

[93]Note in particular that $\Delta_{R}^{conv\,(in)}$ and $\Delta_{R}^{conv\,(in)}$ are determined by R's (first-order) payoffs in Γ, not the underlying fitness payoffs, since the former characterise her incentives.

[94]This is automatically true of type R, given (i) the notational convention I have used that in an ideological game a type θ is associated with both meta- and first-order payoffs, and (ii) the latter have already been specified in Γ.

tives to retain can similarly be calculated as $\Delta^{ret\,(in)}_{M',R} = u_{M'}(\mathbf{z}^*(M',M')) - u_{M'}(\mathbf{z}^*(M',R)) = 0$ and $\Delta^{ret\,(in)}_{R,M'} = u_{M'}(\mathbf{z}^*(R,M')) - u_{M'}(\mathbf{z}^*(M',M')) = 0$. Turning to the ideological incentives to convert and to retain, the particular numerical values are somewhat arbitrary and selected for illustration. The relevant property is that for type-$M^{'}$ players, the incentive to convert is stronger than for type-R players.[95]

Suppose first that $c = 7$. In this case, no players invest and the benchmark result obtains. Suppose instead that $c = 5.5$. In this case, investment only takes place by type M' when matched with type R. Expected fitness for M' is simply $-5.5\varepsilon_R$, the expected cost from investing only when encountering type-R players. Expected fitness for R is $+3\varepsilon_R - \varepsilon_{M'}$, where the second term arises from the fact that R is converted, as discussed above. Equating these two expressions and substituting $\varepsilon_R \equiv 1 - \varepsilon_M$ yields a steady state (i.e. a type distribution where the two types have equal fitness) of $(\varepsilon^*_R, \varepsilon^*_{M'}) = (\frac{2}{19}, \frac{17}{19})$. If type R has a lower population share than $\frac{2}{19}$, then materialists enjoy higher fitness than reciprocators.

Allowing for investment in converting one's opponent's type, and allowing for M' to be ideological thus allows for M' to maintain a stable population share in the long run. Evolutionary pressure, in some form – possibly cultural, as discussed in Chapter 1 of this thesis – may thus explain how and why ideologies sustain themselves. If ideologies themselves can be the object of selection pressure, this suggests a further possibility: that other ideologies may appear (or mutate, in the evolutionary model) based on existing pragmatic types; there may an ideological "arms race", in this sense. To illustrate this possibility in the current setting, suppose now that R is replaced by an ideological reciprocator type denoted $R^{'}$, with $\Delta^{conv\,(id)}_{R'} = 6$, $\Delta^{conv\,(in)}_{R'} = 5$, $\Delta^{ret\,(id)}_{R',M'} = 6$, $\Delta^{ret\,(in)}_{R',M'} = 0$.

Set $c = 5.5$, as before. Note that if $M^{'}$ and $R^{'}$ meet, both invest and so they retain their types. Expected fitness for $M^{'}$ is now $-5.5\epsilon_{R'}$, whereas expected fitness for $R^{'}$ is $+3\varepsilon_{R'} - 5.5\varepsilon_{M'}$. This time there is a unique steady state at $(\varepsilon^*_{R'}, \varepsilon^*_{M'}) = (\frac{11}{28}, \frac{17}{28})$. If type $R^{'}$ has a population share of less than $\frac{11}{28}$, then reciprocators have lower fitness than materialists; if on

[95] As strong ideology, like weak ideology, requires that (3.24) be satisfied, I set $\Delta^{ret\,(id)}_{M',R} > 0$, though this does not affect the subsequent results.

the other hand $\varepsilon_{R'} > \frac{11}{28}$, then the reverse is true. If both types are ideological, then ideology is especially damaging to its adherents when they meet opponents, as they engage in a "war of attrition". If one type is a small enough part of the population, then a relatively large share of their encounters will be with other-type opponents, making them fare worse than the other type.