Selfishness, altruism and utility in everyday two-person random interactions: Effects of strong reciprocity, the common good and the costs of competition

Nipun Agarwal

Centre for Strategic Economic Studies, Business and Law, Victoria University

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Business Administration

December, 2011

Approved by _____
Chairperson of Supervisory Committee

_____

_____

_____

Program Authorized
to Offer Degree_____

Date _____

ABSTRACT

Why should we model two-person random interactions in everyday human activity? If we had to how would we model such interactions? This thesis tries to understand how individuals will behave in 2-person everyday interactions as such interactions comprise a substantial part of our everyday lives. It starts to answers these questions by reviewing Adam Smith's concepts of selfishness (self-interest) and altruism (benevolence) in his books, An Inquiry into the Nature and Causes of the Wealth of Nations (1776) and The Theory of Moral Sentiments (1790) to put this research into context. This literature review is extended to strong reciprocity that develops on the ideas of selfishness and altruism and explains how punishment can impact such behaviour. Game theory and complex system are used to develop the 2-person random interaction model (2PRIM) in order to explain the evolutionary dynamics of selfishness, altruism and strong reciprocity within such interactions. Previous two-person models have not simulated random interactions with strong reciprocity. Results show that selfishness increases rapidly in such interactions and punishment has little or no impact in such situations. However, an increase in the return on selfishness, common or public good or cost of competition in 2PRIM can have a similar impact as strong reciprocity in such two-person random interactions.

STATEMENT OF ORIGINALITY

I declare that the work presented in this thesis is to the best of my knowledge and belief, original and my own work. I also confirm that the material has not been submitted, either in whole or in part, for a degree at this or any other university, or for publication prior to candidature.

Nipun Agarwal	Dr James Doughney
DBA Candidate	Principal Supervisor

ACKNOWLEDGEMENTS

I would especially like to thank my supervisors, Dr James Doughney and Dr Mark Bowden, for their outstanding advice and support. I really appreciate their guidance through my candidature and they have taught me many excellent things through my time with them. I would also like to thank Kate O'Rouke for her valuable feedback and for taking the extra effort of reviewing my research proposal and parts of my thesis and Dr Nick Fredman for editorial suggestions. I would also like to thank Professor Charles Telly, School of Business, Department of Business Administration, State University of New York at Fredonia, for his excellent advice when I was preparing my research proposal. Finally, I would really like to thank my wife, Ashima, for her tolerance and support while I have been writing this thesis. I would also like to thank my father, who has sacrificed so much for me throughout my life.

CONFERENCE PRESENTATION

Chapter 5 was presented at the *15th International Conference in Computing in Finance and Economics*:

Agarwal, Nipun (2009), 'Is it better to be selfish or altruistic in two-person human interactions?', *15th International Conference in Computing in Finance and Economics*, University of Technology, Sydney, 16 July 2009.

TABLE OF CONTENTS

1. INTRODUCTION 10
1.1. INTRODUCTION 10
1.2. THE THEORETICAL APPROACH OF THIS THESIS 10
1.3. THE SIGNIFICANCE OF THE QUESTIONS THE THESIS SEEKS TO ANSWER 13
1.4. AN ORIGINAL CONTRIBUTION TO KNOWLEDGE 15
1.5. OUTLINE OF THE THESIS' ARGUMENT 17

2. SELFISHNESS AND ALTRUISM 23
2.1. INTRODUCTION 23
2.2. ADAM SMITH – THE CROSSROADS OF ECONOMICS AND ETHICS 24
2.3. THE IMPARTIAL AND WELL-INFORMED SPECTATOR AND THE LIMITS OF SELF-INTEREST 30
2.4. SOCIO-BIOLOGY, ECONOMICS AND BEHAVIOURAL ECONOMICS ON SELFISHNESS AND ALTRUISM 33
2.5. NEUROECONOMIC INSIGHTS INTO ALTRUISM AND SELFISHNESS 40
2.6. DEFINING SELFISHNESS AND ALTRUISM 42
2.7. CONCLUSION 43

3. UNDERSTANDING STRONG RECIPROCITY 44
3.1. INTRODUCTION 44
3.2. WHAT IS STRONG RECIPROCITY? 44
3.3. HOW DOES STRONG RECIPROCITY RELATE TO SELFISHNESS AND ALTRUISM? 47
3.4. HOW DO GENETIC CO-EVOLUTION, CO-OPERATION AND OTHER FACTORS AFFECT STRONG RECIPROCITY? 51
3.5. IS STRONG RECIPROCITY AFFECTED BY FAIRNESS, INTENTIONS AND SOCIAL PREFERENCES? 59
3.6. REVIEWING STRONG RECIPROCITY THROUGH NEUROECONOMICS 66
3.7. CONCLUSION 67

4. GAME THEORY AND COMPLEX SYSTEMS 69
4.1. INTRODUCTION 69
4.2. GAME THEORY 70
4.3. COMPLEX SYSTEMS APPLICATION TO ECONOMICS 75
4.4. COMPLEX SYSTEMS AND STRONG RECIPROCITY 77
4.5. EXPLAINING THE STRONG RECIPROCITY MODEL 79
4.6. WORK IN THE TRADITION OF BOWLES AND GINTIS 82
4.7. CONCLUSION 91

5. ANALYSING SELFISHNESS AND ALTRUISM IN TWO-PERSON RANDOM INTERACTIONS 92
5.1. INTRODUCTION 92
5.2. FROM ADAM SMITH TO GAME THEORY AND THE 2PRIM MODEL 96
5.3. THE TWO-PERSON RANDOM INTERACTION MODEL (2PRIM) IN DETAIL 106
5.4. ANALYSING THE SELFISH FACTOR AND ADAM SMITH'S IMPARTIAL SPECTATOR (GUILT) 119
5.5. CHANGES IN THE COST OF COMPETITION AND LEVEL OF SELFISHNESS 124
5.6. COST OF COMPETITION AND MEAN UTILITY 130
5.7. CONCLUSION 134

6. STRONG RECIPROCITY IN TWO-PERSON RANDOM INTERACTIONS 138
6.1. INTRODUCTION 138
6.2. STRONG RECIPROCITY AND THE 2PRIM MODEL 140
6.3. TWO-PERSON RANDOM INTERACTION MODEL (2PRIM) WITH STRONG RECIPROCITY 144

6.4. RESULTS FROM THIS TWO-PERSON RANDOM INTERACTION MODEL (2PRIM) WITH STRONG RECIPROCITY 150
6.5. CONCLUSION 160

7. SUMMARY, LIMITATIONS AND SUGGESTED EXTENSIONS 161
7.1. SUMMARY 161
7.2. CONTRIBUTION TO KNOWLEDGE 164
7.3. LIMITATIONS TO THIS RESEARCH 167
7.4. POSSIBLE APPLICATIONS OF THIS RESEARCH 169
7.5. SUGGESTED EXTENSIONS 170

REFERENCES 173

ATTACHMENT A: PROCESS STEPS FOR THE TWO-PERSON RANDOM INTERACTION MODEL (2PRIM) 208

ATTACHMENT B: *MATLAB* CODE OF MODEL (2PRIM) DEVELOPED IN CHAPTERS 5 AND 6 213

TABLE OF FIGURES

5.1. 2PRIM's structure and dimensions for two-person interactions (maxima and minima) 104

5.2. 2PRIM's formulae for utility from two-person interactions (maxima and minima) 105

5.3. Diagrammatic representation of 1 round of the two-person random interaction model (2PRIM) 109

5.4. Level of selfishness and utility in the standard 2PRIM 112

5.5. 2PRIM formulae for utility from interactions for CGfactor = 1.0 115

5.6. 2PRIM formulae for utility from interactions for CGfactor = 1.5 116

5.7. 2PRIM formulae for utility from interactions for CGfactor = 2.0 116

5.8. Level of Selfishness and Utility in 2PRIM at CGfactor = 1.5 118

5.9. Level of Selfishness and Utility in 2PRIM at CGfactor = 2.0 118

5.10. 2PRIM formulae for utility from interactions for SFfactor = 1.5 120

5.11. 2PRIM formulae for utility from interactions for SFfactor = 2.0 121

5.12. Level of Selfishness and Utility in 2PRIM at SFfactor = 1.5 123

5.13. Level of Selfishness and Utility in 2PRIM at SFfactor = 2.0 123

5.14. Level of Selfishness in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.00 126

5.15. Level of Selfishness in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.25 127

5.16. Level of Utility in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.50 129

5.17. Level of Selfishness in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.75 129

5.18. Level of Utility in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.00 132

5.19. Level of Utility in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.25 132

5.20. Level of Utility in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.50 133

5.21. Level of Utility in 2PRIM with different levels of the Selfish (SFfactor) and CGfactor, when Cfactor = 0.75 133

6.1. Representation of 2PRIM A/S and P/NP continua 145

6.2. Diagrammatic representation of 1 round of the two-person random interaction model (2PRIM) 147

6.3. Diagrammatic representation of rounds 1 and 2 of 2PRIM with Strong Reciprocity 149

6.4. Level of Utility in 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.00) 152

6.5. Level of Selfishness in 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.00) 152

6.6. Level of Utility in 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.10) 154

6.7. Level of Selfishness in 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.10) 154

6.8. Level of Utility in 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.50) 155

6.9. Level of Selfishness 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.50) 155

6.10. Level of Utility in 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.75) 156

6.11. **Level of Selfishness 2PRIM with Strong Reciprocity (Pfactor = 0.50 and Pcost = 0.75) 156**

CHAPTER 1

Introduction

1.1 Introduction

How are we to model the individual, social and evolutionary consequences of everyday two-person random interactions? What effects do selfishness, altruism, reciprocity, competition and the common or public good have in modelling those consequences?

This thesis aims to contribute towards a growing number of interdisciplinary efforts designed to answer such questions. How the thesis proposes to make that contribution is summarised in this introductory chapter. This chapter also outlines important caveats on the thesis' reach. First, however, this introductory chapter will explain precisely where the theoretical approach of the thesis fits, what questions it seeks to answer, why its subject matter represents a significant contribution to knowledge and in which respects its contribution is original.

The particular contribution of the thesis is its focus on *modelling* consequences of everyday two-person random interactions in terms of selfishness, altruism, reciprocity, competition and the common or public good. While it discusses these concepts broadly, the thesis does not draw conclusions about the facts of human nature, evolution or actual behaviour. Rather it models those consequences deductively.

1.2 The theoretical approach of this thesis

Human interaction is the binding thread of our social fabric. While there are different types of human interactions, a significant kind of human interaction is our contact with

strangers in everyday activities. We interact with people we have not, for all intents and purposes, met before in our lives. Wherever we go, and whatever we do, we only interact with some people once in our lifetimes. However, we have vast numbers of encounters of this kind. It is therefore important to understand the effects of how people behave in such interactions. What are the consequences if people behave selfishly or altruistically? What are the individual and social costs and benefits of being selfish or altruistic in such everyday two-person random interactions?

This thesis develops a model to help us to understand such interactions and to find out what happens in defined circumstances if people act either selfishly or altruistically. As Gintis et al. (2005, p. 2-7) explain from the perspective of the social sciences, theories that explore the interactions of selfishness, altruism and reciprocity have become increasingly important in 21st century interdisciplinary research, especially in economics, evolutionary biology and psychology. They have 'become part of a general movement towards transdisciplinary research based on the analysis of controlled experimental studies of human behavior' (Gintis et al. 2005, p. 6) in laboratory and field projects in a variety of settings, in both economically advanced and traditional societies. This research explores the perennial questions that arise from the evidence of everyday experience, in which we observe overt selfishness but also great acts of self-sacrifice. One such question is whether altruism really exists or is merely disguised, enlightened or selfishness with a longer-term perspective. From the perspective of the evolution of human behaviour, Wilson et al. (2009) make strong claims for an answer arising from interdisciplinary research across economics and evolutionary disciplines. They note in regard to this question about the relationship between altruism and selfishness (Wilson et al. 2009, pp. 190-1):

> [This] has been asked in all branches of the human behavioral sciences (e.g., social psychology, sociology, political science, economics, anthropology). Evolutionary theory broadens the scope by examining the evolution of altruism and selfishness in all species ... The same theoretical framework can be used to study human altruism and selfishness. It can even go beyond the study of human genetic evolution to include faster processes of human behavioral change.

In this respect, Wilson et al. (2009) argue for the usefulness of behavioural game theory, as utilised in experimental economics. Gintis et al. (2003) argue that such theory provides a fruitful alternative to traditional or rational-choice economic theory, which is unable to explain co-operative behaviour. In fact the experimental behavioural game-theoretic approach aligns with Adam Smith's view in The Theory of Moral Sentiments (1790), in which Smith suggests that people do feel concerned about other people's welfare in addition to their own. As Gintis (2009, p. 198) puts it:

> Theoretically, experimental economics is already converging with evolutionary theory in two respects. First, human social preferences require a deep explanation in terms of genetic evolution, even if they are highly variable in their individual and cultural expression. Second, fast paced processes of behavioral change count as evolutionary to the extent that they cause the most successful behavioral strategies to increase in frequency over time. There is a generalized replicator dynamic that includes, but goes beyond, genetic evolution (Bowles 2003, Gintis 2000c). A growing number of experimental economists have thoroughly incorporated both perspectives into their own thinking.

This thesis fits into this emerging tradition of interdisciplinary or transdisciplinary research. In particular it draws on the work modelling selfishness, altruism and (strong)

reciprocity championed by Bowles and Gintis (2000), Boyd et al. (2003) and Fehr and Gächter (2002) and, with original variations on the theme, by Eldakar et al. (2007) and Eldakar and Wilson (2008).

1.3 The significance of the questions the thesis seeks to answer

The significance of this thesis' subject matter derives from the questions it seeks to answer, which are:

1. How might we model individual, social and evolutionary consequences of everyday two-person random interactions?
2. What effects might selfishness, altruism, reciprocity, competition and the common or public good have when they are included in a model that simulates those consequences?

That is, it asks that if we model selfishness, altruism, reciprocity, competition and the common or public good in certain ways, what might be the individual, social and evolutionary consequences of everyday two-person random interactions? Developing models to assist those exploring the deeper questions of human nature, evolution and actual behaviour is a significant undertaking.

For example, those who argue the case that altruism and co-operation are an evolved feature of human behaviour and that they are not just forms of enlightened or long-run self interest rely heavily of the role of reciprocity in random (or 'one-shot', 'non-repeated') interactions (or 'games'). Indeed, our 'ability to interact peacefully in one-shot interactions with strangers may prove to be one of the most remarkable traits

of our ... species' (Silk 2005, p. 64). Central to this trait is strong reciprocity, or our willingness to sanction selfish behaviour even if we do so at a cost. Silk explains in a study of the evolution of co-operation among primates that concept of 'strong reciprocity emerged from carefully designed experimental studies on humans that revealed surprisingly high levels of altruism in one-shot interactions with strangers' (2005, p. 63). She adds that comparable evidence regarding the treatment of strangers among non-human primates would be hard to find.

In this discussion, the role of random interaction (i.e. interaction among strangers) is key. If interaction were repeated, it would be an easy matter to reconstruct altruism as a matter of self-interest. That is, if we knew we were to interact with someone repeatedly, it would pay us not to be selfish lest we be treated in the same way. Therefore, the altruism-as-selfishness argument goes, if altruistic behaviour arises in one-shot experiments or games it is just because people misapply behaviour that they have learned from their experience of repeated interactions. If they were rational they would opt for selfishness because they could get away with it, as they never have to interact again. Gintis (2000c, p. 262) disagrees, referring specifically to the evolutionary models of Trivers (1971) and Axelrod and Hamilton (1981) and, implicitly, to selfish-gene theories such as that of Dawkins (1976):

> However, as we know from the theory of repeated games ... , reciprocity in the above sense is just 'enlightened self-interest,' ... In effect, Trivers's reciprocal altruist and the Axelrod-Hamilton Tit-for-Tat behave little differently from *Homo economicus* ... The *Homo reciprocans* who emerges from laboratory experiments, by contrast, provides a more robust basis for pro-social behaviour, since he does not depend upon frequently repeated interactions to induce him to cooperate and punish defectors.

Homo reciprocans demonstrate a behaviour labelled as *strong reciprocity*, which means that they co-operate with others and punish those who violate social norms even at a personal cost to themselves and when they are no rewards from this behaviour.

Critics suggest that reciprocal behaviour in one-shot games is just a confused carryover of the subject's extensive experience with repeated games in everyday life to the rare experience of the one-shot game in the laboratory. This is incorrect. Human beings in contemporary society are engaged in one-off games quite often, instead any interaction that we have with a stranger could fall into this category. Also, when people face major events in their lives, for example, fight on the frontline in a war or experiencing a natural disaster, these can be classified as one-off experiences where people exhibit strong reciprocity similar to what you would notice in a laboratory experiment. Moreover, *the fact that humans often 'confuse' one-shots and repeated interactions, when they clearly have the cognitive mechanisms to distinguish, suggests that the 'confusion' may be fitness-enhancing*. (original emphasis; Gintis et al. 2005, pp. 8, 25-6, 2008, pp. 242-3, 249-52).

1.4 An original contribution to knowledge

Thus far I have taken some space, and quoted extensively, to establish the point that to try to study and model random, one-shot or non-repeated (random) interactions represents a significantly useful area of research. However, is the particular contribution envisaged by this thesis original? That is, do the models it develops represent an original contribution to knowledge in this emerging interdisciplinary field? The answer, again, arises from the research questions put forward in the previous section.

The first question specifies a particular kind of everyday random interaction, those involving two people only (or 'dyadic'). This thesis develops a computer simulation model, the two-person random interaction model (2PRIM), to assist in understanding

such interactions. To evaluate the respective effects of selfishness, altruism and strong reciprocity on the consequences of two-person random interactions, the model allows for those behaviours or traits to be represented randomly along a continuum. While other scholars address similar random or one-shot interactions, none seem to combine these three aspects in modelled interactions, i.e. as two-person, as continuous-trait and as involving strong reciprocity. The closest parallels are those of Eldakar et al. (2007) and Eldakar and Wilson (2008). The former models strong reciprocity in N-person interactions, with traits allocated 'that initially vary uniformly ... between 0 and 1 at 0.1 increments' (Eldakar et al. 2007, p. 199). However, N is always greater than two. The latter study allocates traits or behaviours for selfishness, altruism and strong reciprocity as 'pure strategies', which is to say that members of a group of N are either pure altruists or purely selfish and either strong reciprocators or not (Eldakar and Wilson 2008, p. 6982). Moreover earlier work on co-operation using two-person models relies on 'non-random interactions or guarded cooperation' (Eldakar and Wilson 2008, p. 6982, citing Axelrod 1984, Hamilton 1964, 1975, Axelrod and Hamilton 1981, Maynard Smith 1982) and does not consider strong reciprocity.

The first contribution of this thesis therefore is that it models a *significant* problem that scholars seek to understand, namely random, one-shot or non-repeated human interactions. This problem is one we face in our everyday lives in our communities and, if we travel, beyond them. We regularly engage in random interactions in our working, professional and leisure activities. As Silk (2005, pp. 63-4) has pointed out, such interactions are of evolutionary significance.

Secondly, the thesis contributes originally to knowledge by modelling a specific set of random interactions, namely those involving two persons, continuous-trait attributes

and strong reciprocity. Thirdly, the thesis creates an original game-theoretic computational model (2PRIM) in order to contribute to our understanding of the individual, social and evolutionary consequences of everyday two-person random interactions in defined circumstances. Fourthly, the thesis contributes originally to knowledge by developing the foregoing contributions within the context of a theoretical discussion of the literature – including the work of Adam Smith – of this emerging multidisciplinary field of research. Indeed it is only within the context of that discussion that the questions posed by this thesis can be understood. It is only within that context, too, that the models presented in this thesis make sense. The final contribution of this thesis is that the 2PRIM shows that pay-offs for players can be modified through changes in the return on selfishness, common good or the cost of competition, which can be a substitute to applying strong reciprocity within random two-person interactions.

1.5 Outline of the thesis' argument

The thesis begins a review of the relevant literature in Chapter 2, by, following Wilson et al. (2009, p. 198), returning to Adam Smith's contribution to our understanding of selfishness and altruism. Smith is indeed an appropriate starting point, even though he is one who is often misunderstood as being an advocate of selfishness. Smith's Theory of Moral Sentiments (1790), as the name suggests, grounds the origins of human morality in natural human sentiments, affections, passions or emotions. Moreover he maintains that the 'great division of our affections is into the selfish and the benevolent' (Smith 1790, p. 267). The selfish affections recommend to us, on the one hand, the virtue of prudence, which exemplifies proper self-interest in our own happiness (Smith 1790, p.

262). On the other, however, Smith distinguished between self-love (or self-interest) and selfishness in conduct and character, 'using "selfishness" in a pejorative sense for such self-love as issues in harm or neglect of other people' (Raphael and Macfie 1976, p. 22). Benevolent affections, in contrast, recommend to us the virtues of justice and beneficence, which exemplify concern for the happiness of others. Justice 'restrains us from hurting' others' happiness, and beneficence, the nearest word Smith uses to altruism, 'prompts us to promote that happiness' (Smith 1790, p. 262).

Understanding Smith's work also brings us nearer to a theme that resonates throughout this thesis. This theme is reciprocity, and Smith's views above lead us unerringly to it (for example, Smith 1790, pp. 78-91, 95-6). In a notable passage on the propriety of humans' more turbulent and selfish emotions, in contrast to those emotions of a more benevolent kind, Smith (1790, p. 25) foreshadows developed theory:

> And hence it is, that to feel much for others and little for ourselves, that to restrain our selfish, and to indulge our benevolent affections, constitutes the perfection of human nature; and can alone produce among mankind that harmony of sentiments and passions in which consists their whole grace and propriety. As to love our neighbour as we love ourselves is the great law of Christianity, so it is the great precept of nature to love ourselves only as we love our neighbour, or what comes to the same thing, as our neighbour is capable of loving us.

Chapter 3 of the thesis analyses the concept of strong reciprocity. As noted earlier, the thesis draws on the work modelling strong reciprocity undertaken *inter alia* by Bowles and Gintis (2000), Boyd et al. (2003) and Fehr and Gächter (2002) and, with different features, by Eldakar et al. (2007) and Eldakar and Wilson (2008). Gintis et al. (2003, p. 8) define strong reciprocity as '*a predisposition to cooperate with others, and*

to punish (at personal cost, if necessary) those who violate the norms of cooperation, even when it is implausible that those costs will be recovered at a later date' (original emphasis, Gintis et al. 2008, p. 243, 2000a, p. 262). The chapter explores the concept in detail, reflecting its important place in contemporary behavioural game theory. The literature reviewed in the chapter is also especially important regarding human interactions in the 'non-repeated and anonymous situations' (Gintis et al. 2005, p. 8) that comprise the focus of the questions this thesis endeavours to answer. Chapter 3 covers the literature generally, but it also explains in greater detail the complex relationship between first- and second-order altruism and altruism modelled in the work of Eldakar et al. (2007) and Eldakar and Wilson (2008). Where necessary it contrasts this approach to the more straightforward models of strong reciprocity.

Using insights from evolutionary game theory, complex systems and behavioural economics set out in Chapter 4, the thesis proposes, as an example, a computational model of two-person random human interactions (called the 2PRIM). The 2PRIM has the basic form of a two-person Prisoners' Dilemma game. Where, Prisoners' Dilemma is a classical game theory problem that was introduced by Merrill Flood (1952) and Melvin Dresher (1961). However, it operates according to a computational procedure designed to model many interactions (games) and to embed an evolutionary process. As Chapter 5 explains, two individuals from a pool of 1000 randomly play a Prisoners' Dilemma type of game (called a 'round'). This process repeats for 100,000 rounds (called a 'game'), after which the model calculates individuals' payoffs from interactions. The model eliminates and randomly replaces the 10 per cent of the game's individuals who have the lowest payoffs. The model runs 1000 games, i.e. there are 10^8 rounds.

A standard version of the 2PRIM presented in Chapter 5 demonstrates by design firstly how selfishness might dominate in one-off random interactions with strangers but also secondly how an evolutionarily stable outcome that allows for some level of altruism might emerge. The components of the model are:

1. Attribution of altruistic (A_i) and selfishness (S_i) traits to an individual *i* on a continuum from 0 (purely selfish) to 1 (purely altruistic), where $1 - S_i = A_i$.
2. A common-good factor (*CGfactor*) designed to incorporate the return to individual *i* due to altruistic interaction with individual *j*.
3. A selfishness factor (*Sfactor*) designed to incorporate return to individual *i* of *i's* selfishness at the expense of individual *j*.
4. A cost-of-competition factor (*Cfactor*) designed to incorporate the cost to individual *i* of selfish competition with individual *j*.
5. A payoff or utility function that combines 1-4 and so that the game procedure can eliminate and randomly replace those with the lowest 10 per cent of payoffs in order to model evolutionary dynamics (as discussed above).

The model's outcomes illustrated in Chapter 5 see selfishness increase rapidly in the standard form of the 2PRIM model. However, changes in the common- (or public-) good factor, selfishness factor and cost of competition factor have a significant impact and change the levels of selfishness-altruism and payoff or utility emerging in concert with evolutionary stability.

Chapter 6 then develops the standard model to include the application of probabilistic punishment as a proxy for strong reciprocity. Four additional components add to the standard model:

1. Attribution of punisher (P_i) and non-punisher (NP_i) traits to individual *i* on a continuum from 0 (pure punisher) to 1 (pure non-punisher), where $1 - NP_i = P_i$.
2. A punishment factor (*Pfactor*) designed to incorporate the cost to individual *i* of *i* being punished for selfishness at the expense of individual *j*.
3. A cost-of-punishing factor (*Pcost*) designed to incorporate the cost to individual *i* of punishing individual *j* for selfishness.
4. A modification of the payoff function and procedure at 5 in order to model the effects of altruistic punishment.

The results presented in Chapter 6 show the effects of introducing punishment into the model. Results from Chapter 6 provide a significant contribution, as they differ from the way punishment has been applied in previous models. Bowles and Gintis (2000), Boyd et al. (2003) and Fehr and Gächter (2002) have considered strong reciprocity, while Eldakar et al. (2007) and Eldakar and Wilson (2008) have considered selfish punishment. However, as often happens in real life, when one finds someone being more selfish than themselves, even while one is altruistic, one will temporarily become selfish to teach the more selfish person a lesson. Results in Chapter 6 show how the dynamics of the game changes when the less selfish player becomes temporarily more selfish. Effectively, this raises the number of altruists that survive compared to when there was no punishment involved. Resultantly, this increases the level of altruism in the pool and effectively the level of selfishness decreases. Changes in the cost of

punishment, selfishness, altruism, the common good and cost of competition factors also affect this equilibrium. Chapters 5 and 6 also set the developments in the model into the particular contexts of the relevant literature, explaining where necessary differences and similarities in approach. In this way the original aspects of the contribution of this thesis to the modelling exercise will be apparent to the reader.

Chapter 7 summarises the contribution of the thesis and explains its limitations. While this concluding chapter will reinforce the point, it is important also for the reader to understand these limitations at the outset. That is, in order to avoid misinterpretation of purpose, it is necessary to understand what it is that the thesis is not trying to accomplish and, indeed, cannot accomplish. In particular, it is important readers do not think that the thesis intends to draw substantive conclusions concerning human selfishness, altruism, reciprocity, competition and the common or public good. Nor does it say how these aspects of human nature, evolution and behaviour have actually manifested themselves in reality.

Instead the thesis has more modest objectives. Based on a comprehensive discussion of the above concepts and the literature surrounding them, the purpose of the thesis is to offer a modelling approach that those who are endeavouring to answer the big questions can use as a tool. The modelling approach the thesis contributes is *therefore* deductive in structure. That is, its asks, if human selfishness, altruism, reciprocity, competition and the common or public good were found to be such-and-such and related to each other with such-and-such weights attached to them, what might be the individual, social and evolutionary consequences of everyday two-person random interactions?

CHAPTER TWO

Selfishness and altruism

2.1 Introduction

This chapter defines and discusses the concepts of selfishness and altruism. It will start with Adam Smith, whose name and work appear in many of the works cited already in the Introduction (for example Gintis et al. 2005, Wilson et al. 2009). Selfishness, or assumed correlates such as self-interest, self-love, self-regard or prudence, is a concept that is prevalent in Adam Smith's book *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776). On the other hand, a concept that is the opposite of selfishness, namely altruism, or in Smith's terms benevolence and beneficence, features prominently in Adam Smith's 'other' book, *The Theory of Moral Sentiments* (1790).

Firstly, the chapter will focus on Smith and the insights that we can still fruitfully extract from his seminal work. To avoid delving into the many disputes over the interpretation of Smith's work, the first section of this chapter relies on Smith's own voice. Secondly, the chapter discusses selfishness and altruism from the viewpoints of moral philosophy, socio-biology, behavioural economics and neuroeconomics. The development of evolutionary game theory has been closely related to these fields. Finally, this discussion will help to define the concepts of altruism and selfishness for use throughout the thesis and also to lead into the discussion in Chapter 3 of the relationships between self-interest, altruism and a recently developed concept called strong reciprocity. The latter involves the interaction of self-interest, altruism and punishment.

2.2 Adam Smith – the crossroads of economics and ethics

Adam Smith published the first edition of *The Theory of Moral Sentiments* in 1759. A sixth edition, 'published shortly before Smith's death in 1790, contain[ed] very extensive additions and other significant changes' (Raphael and Macfie 1976, p. 2). He published *An Inquiry into the Nature and Causes of the Wealth of Nations* in 1776. Both are seminal works in (classical) economics and ethics and warrant study in their own right, and not merely for Smith's latter-day reputation as the father of free-market economics.

In *The Theory of Moral Sentiments* Smith does not use the word altruism. Rather, he uses the concepts of benevolence and beneficence for what, today, we might call expressions of altruism or kindness, or the unselfish act of giving or helping other humans without expecting anything in return.

Before discussing Smith, it is worthwhile distinguishing between his two relevant terms. According to the *Concise Oxford Dictionary* (1976), the adjective benevolent means to be 'desirous of doing good' or to be 'charitable' or 'kind and helpful'. The adjective beneficent, however, means 'doing good' or '(showing) active kindness'. The difference, slight as it is, concerns a person either having the sentiment, emotion, affection or passion to be kind, helpful and charitable (benevolence) or to *be* kind, helpful and charitable in action (beneficence).

Smith contrasted benevolent and selfish affections. He maintained that the 'great division of our affections is into the selfish and the benevolent' (Smith 1790, p. 267). Yet he saw each of them as motivating different aspects of a virtuous character and virtuous conduct. As he wrote 'of Virtue':

> Concern for our own happiness recommends to us the virtue of prudence: concern for that of other people, the virtues of justice and beneficence; of which, the one restrains us from hurting, the other prompts us to promote that happiness. Independent of any regard either to what are, or to what ought to be, or to what upon a certain condition would be, the sentiments of other people, the first of those three virtues is originally recommended to us by our selfish, the other two by our benevolent affections. Regard to the sentiments of other people, however, comes afterwards both to enforce and to direct the practice of all those virtues; and no man during, either the whole of his life, or that of any considerable part of it, ever trod steadily and uniformly in the paths of prudence, of justice, or of proper beneficence, whose conduct was not principally directed by a regard to the sentiments of the supposed impartial spectator, of the great inmate of the breast, the great judge and arbiter of conduct (Smith 1790, p. 262).

While our selfish affections recommend to us the virtue of prudence, which exemplifies proper self-interest in our own happiness, they also can overcome the 'feeble spark of benevolence which nature has lighted up in the human heart' (Smith 1790, p. 137) and cause improper expressions of self-love (or self-interest) and selfishness in conduct and character. In such instances, Smith uses '... "selfishness" in a pejorative sense for such self-love as issues in harm or neglect of other people' (Raphael and Macfie 1976, p. 22, see also Smith 1790, pp. 9-10, 135-8, 317). This is why we require the virtue of self-command, a virtue that will bring down the temper of all of our passions, but especially the 'turbulent and mutinous' and selfish (Smith 1790, pp. 247, 250, 266), to the level that we ourselves think, on sober reflection, are proper:

> The man who acts according to the rules of perfect prudence, of strict justice, and of proper benevolence, may be said to be perfectly virtuous. But the most perfect knowledge of those

> rules will not alone enable him to act in this manner: his own passions are very apt to mislead him; sometimes to drive him and sometimes to seduce him to violate all the rules which he himself, in all his sober and cool hours, approves of. The most perfect knowledge, if it is not supported by the most perfect self-command, will not always enable him to do his duty (Smith 1790, p. 247).

While the benevolent and selfish sentiments recommend the other virtues, self-command is 'upon most occasions, principally and almost entirely recommended to us by ... the sense of propriety, by regard to the sentiments of the supposed impartial spectator'. If it were not for 'the restraint which this principle imposes, every passion would, upon most occasions, rush headlong, if I may say so, to its own gratification' (Smith 1790, p. 262).

How does this balanced approach fit with the contrary view most would hold of *The Wealth of Nations*? In the latter work, Smith seemingly presents a divergent ideology to that of *The Theory of Moral Sentiments*. In *The Wealth of Nations* he emphasises the concept of self-interest and states, in a well-rehearsed quotation drawn from his discussion of the division of labour, that people are more willing to help when it is in their own self-interest rather through benevolence (Smith 1776, pp. 117-9). Smith notes how individuals need the co-operation of 'multitudes' but scarcely have the time to make a few friends. Hence, it 'is in vain' for someone to expect 'the help of his brethren ... from their benevolence only'. Instead someone 'will be far more likely to prevail' if she or he can win the favour of others by appealing to their self-love, or their own advantage. This fits with a natural human 'propensity to truck, barter and exchange one thing for another'. Anyone who 'offers another a bargain of any kind' suggests mutual advantage:

> Give me that which I want, and you shall have this which you want, is the meaning of every such offer; and it is in this manner that we obtain from one another the far greater part of those good offices which we stand in need of. It is not from the benevolence of the butcher the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity, but to their self-love, and never talk to them of our own necessities, but of their advantages (Smith 1776, pp. 118-9).

In contrast, *The Theory of Moral Sentiments* opens by explaining altruistic motivation in terms of the manner in which humans can empathise with other people's sorrow and how the most selfish person can also feel the pain and sorrow felt by another human being. This quotation is also becoming well-rehearsed:

> How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it. Of this kind is pity or compassion, the emotion which we feel for the misery of others, when we either see it, or are made to conceive it in a very lively manner. That we often derive sorrow from the sorrow of others, is a matter of fact too obvious to require any instances to prove it; for this sentiment, like all the other original passions of human nature, is by no means confined to the virtuous and humane, though they perhaps may feel it with the most exquisite sensibility. The greatest ruffian, the most hardened violator of the laws of society, is not altogether without it (Smith 1790, p. 8).

Among others, Daniels (1898), Bittermann (1940), Robbins (1952), Macfie (1959), Cropsey (1977) and Campbell (1967) provide an understanding of Adam Smith's ideas on self-interest, individualistic passions, benevolence and the impartial spectator. They

state that Adam Smith's theory is based on the fact that people have their own interest at the forefront and society is the product of these individualistic interests. While human passions are part of human character and personality, these passions and sentiments evolve with changes in social relationships (Lamb 1974, p. 671). These changes should therefore create a dynamic societal environment with a competition between self-interested and cooperative impulses. Lamb comments that, 'Fundamentally these two methods, the individualist and the social, are two ways of looking at the same social development' (Lamb 1974, p. 672), restating the 'great division of our affections is into the selfish and the benevolent' as proposed by Adam Smith and arguing that these two qualities should possibly be viewed complimentarily.

According to Morrow (1969, pp. 166-7), 'There is a unity of spirit and aim in Adam Smith's treatment of these separate divisions of moral philosophy (i.e. ethics, political economy, and jurisprudence) that cannot be doubted'. Lamb (1974, p. 673) also adds that, in *The Theory of Moral Sentiments*, 'Smith recommends "benevolence" for universal human improvement' but in *The Wealth of Nations* 'recommends "self-interest" not only as the most powerful motive for human improvement but as the motive which should actuate men in their economic relations'. This point is further clarified by Morrow (1969, p. 159):

> In fact, if those who believe there was a discrepancy between *The Theory of Moral Sentiments* and the *Wealth of Nations* had but taken the pains to consult the former work thoroughly, a great deal of this alleged discrepancy would have disappeared. It is true that in *Moral Sentiments* Adam Smith opposed the egoistic doctrine that man acts only from self-love, and exalts benevolence as the highest virtue. There are other inferior virtues recognized, such as prudence, frugality, industry, self-justice, but when so regulated they are conducive to the welfare of the general public as well as of the individual. The

important consideration is that these self-interested activities must be regulated by justice. In short, unregulated self-interest is no more advocated in the *Wealth of Nations* than it is in the *Moral Sentiments*, whereas in the latter work the moral value of the inferior virtues, when properly regulated, is fully recognized.

Ashraf et al. (2005, p. 2) have also denied any inconsistencies between Adam Smith's books. They state that an individual as defined in *The Theory of Moral Sentiments* needs to consider the concepts of human emotion (including self-interest) against the impartial spectator (i.e. a third person reviewing the situation from an objective viewpoint). They felt that humans also show compassion for others and are concerned with the concepts of fairness and justice in society. Montes (2003) and Evensky (2005) support a similar understanding of Adam Smith's work. In the same vein, Klein (2003) discusses how neo-Darwinists such as Frans de Waal have shown experimentally through behavioural research that monkeys and apes exhibit a blend of competition and co-operation. In contrast to narrow Social Darwinist 'survival of the fittest' perspectives, 'the findings of the neo-Darwinists like de Waal support Smith's view that in nature competition takes place within a cooperative context – or, at least, that they are interconnected'. That is, 'they maintain that moral activity, rather than destructive dog-eat-dog competition, is necessary to achieve the goals of natural selection' (Klein, p. 387).

Despite the weight of scholarly opinion to the contrary, a narrow interpretation of Smith's concept of self-love, supposedly derived from in *The Wealth of Nations*, became the fundamental assumption in economics. This assumption is that humans are a species of *homo economicus* and are merely utility-maximising, self-interested individuals. Yet, as Raphael and Macfie (1976, p. 29) explain, it is surely a mistake to

contrast *The Wealth of Nations* (WN) and *The Theory of Moral Sentiments* (TMS) and thereby to take such a one-sided view of Smith.

> Of course WN is narrower in scope and far more extensive in the working out of details than is TMS. It is largely ... about economic activity and so, when it refers to motivation, concentrates on self-interest. There is nothing surprising in Adam Smith's well known statement (WN I.ii.2): 'It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest.' Who would suppose this to imply that Adam Smith had come to disbelieve in the very existence or the moral value of benevolence? Nobody with any sense (Raphael and Macfie 1976, p. 29).

2.3 The impartial and well-informed spectator and the limits of self-interest

Adam Smith believed in a limited role only for regulation by the state in economic affairs. He felt that market psychology, associated with self-interest, coupled with proper benevolence and justice, would provide the necessary restraint. Nevertheless, according to Nobel economics laureate Amartya Sen (1993, p. 45):

> The importance of business ethics is not contradicted in any way by Adam Smith's pointer to the fact that our regards to our own interests provide adequate motivation for exchange. There are many important economic relationships other than exchange, such as the institution of production and arrangements of distribution. Here business ethics can play a major part. Even as far as exchange is concerned, business ethics can be crucially important in terms of organization and behaviour, going well beyond basic motivation.

Within the broader social context Sen's point is all the more apposite. Indeed for Smith the sense of justice plays an especially important role. Its purpose is to ensure that individuals would not hurt, harm or injure others, either in their person or their interests. The sense of justice also should prevent us from diminishing others' happiness in pursuit of our own (Smith 1790, p. 196). Hence, for Smith, a system of justice should prevail in society to provide for individual security, fairness and integrity, and the state was a body that could limit self-interest where it tends to go out of control (Suttle 1987). The rules of justice are strict and exact, and violations require punishment.

The idea that self-interest could express itself improperly was reflected in the manner in which we noticed Smith distinguish proper self-love (or self-interest) from harm-causing selfishness (Raphael and Macfie 1976, p. 22). What is more, the limits to self-interest implied by justice also applied to competitive commerce. While people will commonly 'indulge' an individual's self-love:

> … so far as to allow him to be more anxious about, and to pursue with more earnest assiduity, his own happiness than that of any other person … In the race for wealth, and honours, and preferments, he may run as hard as he can, and strain every nerve and every muscle, in order to outstrip all his competitors. But if he should jostle, or throw down any of them, the indulgence of the spectators is entirely at an end. It is a violation of fair play, which they cannot admit of. This man is to them, in every respect, as good as he: they do not enter into that self–love by which he prefers himself so much to this other… (Smith 1790, pp. 126-7).

Coase (1976), Gramm (1989) and Ashraf et al. (2005) believe that Smith therefore argued the need for a balance between the selfish passions and the judgement of the

impartial spectator (Smith 1790, pp. 135-6): a balance that is sustained by liberty, fairness and justice. While the impartial spectator governed the appropriateness of all of our sentiments (including the benevolent), he had a special role regarding both the unsocial passions, such as resentment, hatred and anger (Smith 1790, pp. 85-8) and excessive self-love. The social passions, such as generosity, humanity, kindness, compassion and mutual friendship (Smith 1790, p. 38, see also Kennett 1980a and 1980b), were less apt to go awry.

Smith explained that within each person there exists both a partial spectator and an impartial spectator who compete for prevalence (Smith 1790, p. 181). However, it is the latter – the 'impartial and well-informed spectator … the man within the breast' – who must be 'the great judge and arbiter' of right and wrong, justice and injustice (Smith 1790, pp. 161-2, 262). It is the hypothetical impartial spectator who 'holds us to account' for any 'omissions or violations' of virtuous character and conduct by the exercise of the awesome (awful) virtue of self-command (Smith 1790, pp. 262-3). An individual can judge their own actions if they view their own actions from the viewpoint of an impartial spectator.

Smith felt that, complementary to feelings of self-interest and individualism, the proper expression of which gives the virtue of prudence, the virtues of proper beneficence and justice perfected the natural feelings or affections of benevolence. Regardless of however self-interested humans are, they will nonetheless still have an interest in helping other people around them. Though we do so imperfectly, he also explained, we can even feel what another person is feeling to the extent that we can imagine the other person's pain (Smith 1790, p. 9). Individuals have natural sense of connection, beginning with kinship, and this thread weaves the social fabric. The virtue of beneficence even recommends to us an order in which we should care for others and

even societies, from those closest to us by kin, friendship, neighbourhood, country and by mutual engagement to those less so connected (Smith 1790, pp. 232-44). Such relationship between individuals creates the grounds for altruism. Smith (1790, p. 245) even goes so far as to say:

> Though our effectual good offices can very seldom be extended to any wider society than that of our own country; our good-will is circumscribed by no boundary, but may embrace the immensity of the universe. We cannot form the idea of any innocent and sensible being, whose happiness we should not desire, or to whose misery, when distinctly brought home to the imagination, we should not have some degree of aversion.

Such expansive benevolence was possible for Smith because he believed that, through sympathy, individuals can transcend the limits of their own individuality (i.e. self-interest and other individual passions). Nevertheless, while the power of 'sympathy', 'fellow-feeling' or empathy gave us the ability to enter imaginatively into the circumstances of others, the 'impartial spectator', who educates our sentiments towards propriety, merit, duty and virtue, was still critical to fairness in the way we judge ourselves and others. Self-interest had to be controlled and self-command enabled people to restrain their selfish passions helping them focus their efforts toward socially beneficial objectives (Smith 1790, p. 255).

2.4 Socio-biology, economics and behavioural economics on selfishness and altruism

The thesis so far has explored Adam Smith's work regarding selfishness and beneficence, referring to him as being at the crossroads of economics and ethics. In this

tradition, Boulding (1969), Werhane (2000), Dwyer (2005) and Kaplow and Shavell (2007) see economics as a moral science. Others, including Hirshleifer (1977, 1978, 1985), Becker (1976), Kitcher (1998), Koslowski (1999) and Robson (2001), find that socio-biology and psychology when integrated with economics can contribute to a better understanding of Adam Smith's ideas of altruism and self-interest. Subsequent research in the area of socio-biology and behavioural economics has continued to make the connections (for example, Ashraf et al. 2005). Frantz (2000), for example, notes that Adam Smith's use of the concepts of self-interest, intuition, sympathy and the impartial spectator have commonalities with Darwin's and Piaget's work. In this section the thesis will focus less on Smith and more intently on cross-disciplinary contributions to our understanding of altruism and selfishness.

Socio-biology is the scientific study of the biological (especially ecological and evolutionary) aspects of social behaviour in animals and humans (Soanes and Stevenson 2005). It is a particularly helpful discipline for understanding the effects and nature of altruism and selfishness. The eminent biologist E.O. Wilson (2000, p. 578) describes altruism as a 'self-destructive behaviour performed for the benefit of others'. More generally, socio-biologists call a person's behaviour as altruistic if this behaviour will benefit the recipient more than it does the person performing altruistic actions. Margolis (1984, p. 15) states that factors that define altruistic behaviour occur when a person can benefit more from his own actions if they had not considered the effect of their actions on other people. Liebrand (1986) believes, in the context of game theory, that altruists are those who provide greater weight to others' than to their own outcomes in deciding on game strategies.

Psychologists, however, tend to consider both the intentions and cost to the actor when defining altruism (Krebs 1987). In order to define altruism further, Hill (1984),

for example, explains that altruism can exist in the form of: egoism and unselfish altruism. Batson and Shaw (1991), Vine (1983) and Krebs (1987) also differentiate these two types of altruism. Wilson (2000, p. 371) refers to unselfish altruism as 'hard-core altruism'. Simmons and Marine (2002, p. 228) refer to the decision to donate a kidney to a relative as being almost immediate and therefore this can be defined as 'impulsive altruism'. Piliavin et al. (1981, p. 238) note that 'the factors that affect impulsive altruism – clarity, reality and involvement with the victim – have been demonstrated to be related to greater levels of bystander arousal'. Krebs (1987, p. 113) also concludes that:

> Evidence on impulsive altruism suggests that ... humans ... may be genetically disposed to engage in impulsive acts of helping ... The findings that prior experience with a victim facilitates impulsive helping is consistent with evidence on familiarity in support of the possibility that impulsive helping is an anachronistic anomaly.

Bergstrom and Stark (1993) also find that a genetic tendency to co-operate (see Simmons and Marine 2002 for an example) and shared cultural inheritance (i.e. sharing the same values with others) can increase an individual's payoff in socio-biological game-theoretic models and the possibility to survive in the future.

After an extensive survey of models of altruism across multidisciplinary areas (i.e. sociology, economics, biology and psychology), Khalil (2001) comes to the conclusion that four different types exist: egoism, egocentrism, altercentrism and altruism. Egoism is a strategic act of reciprocity in order to enhance future benefits in infinite encounters. Egocentrism is similar to a selfless act, where the actor sacrifices for someone else, but in which the actor gains happiness by seeing the other person's satisfaction from his or her actions. Altruism is an action taken by a person that is deemed not to be strategic or

obligatory and is selfless in its nature. Altercentrism falls between egocentrism and altruism, where altercentric individuals cannot express the voluntary and varied characteristic of altruists (Khalil 2001).

Khalil (2001) believes that Adam Smith's view of a person is that they are able to show sympathy without disconcerting their own self-interest completely. This occurs as, in The Theory of Moral Sentiments, Smith explained that human society was not being run by means of self-interest and self-indulging passions, though prudent pursuit of self-interest was a virtue rather than a vice. Sympathy, the unique feature of Smith's system, could account for both fraternity and, by understanding others' prudential needs, for self-interest. Moreover, there was no requirement in Smith's thinking for an external system of ethics in society because everyday human interaction creates this understanding (Khalil 2001).

Kahana (2005) demonstrates using game theory that altruism can be sustained through evolution even if there are no genetic links between players and altruism increases regardless of the fact that the game is played with individuals that were selected at random. On the other hand, Mueller (1986) feels that co-operation is a learned behaviour and that egoism may be the dominant human trait. That is, it is useful for humans to utilise adaptive egoism – that is, to adapt to other people's views and learn to co-operate – rather than to be purely egoistic. However, Nieli (1986) and Schenk (1987), offer a different view. They believe that a feeling of kinship within the group creates a sense of self-interest against individuals in society who are not part of that group, as each individual within the group tries to protect the interests of the other individuals within the group.

Cooper and Wallace (2004) explain that altruism can succeed in a game if groups are secluded for multiple periods. In such a game it is also important to consider the size

of the group, cost of being altruistic and the number of periods that these groups have been secluded. Additionally, they state that these groups should be of medium size for altruism to flourish as too large or small groups cannot sustain altruism. On the other hand, Hansson and Stuart (1992, p. 301) show a positive connection between socialisation and altruism in their biological experiment. They also find that traditional biological models cannot explain such an equilibrium and their biological model depicts the results from Adam Smith's thinking from *The Theory of Moral Sentiments*. These results are in line with Adam Smith's, within the limits explained in the previous section. Khalil (2004a, 2004b) supports this argument finding a similar connection between socialisation and altruism.

In direct contrast with those who emphasise altruism, the eminent biologist Richard Dawkins (1976), Brewer and Caporael (1990) and the Nobel laureate economist Gary Becker (1976) argue that selfishness is the key factor in evolutionary survival. Becker (1976, 1981, 1993) says that the crossover between economics and socio-biology has provided a better insight into selfish and altruistic behaviour. He believes that it would be useful to combine the analytical skills used in economics with those used in population genetics and socio-biology, as both fields would gain from this interaction (Becker 1976, p. 826). Moreover, attributes including self-interest and altruism towards kin, which economists represent as certain given preferences, could be explained through traits related to genetic fitness in biology (Michael and Becker 1973). Becker (1973, p. 826) explains that survival is in essence the result of utility maximisation in any situation: 'To demonstrate this I have shown how the central problem of socio-biology, the natural selection of altruism, can be resolved by considering the interaction between the utility maximizing behaviour of altruists and egoists'. That is, altruism is a form of utility-maximising, or ultimately selfish, behaviour. In other words, altruism is

disguised selfishness, so an altruist would expect something in return. Therefore, an altruist would only do something for you if he/she can obtain something in return from you at a later stage.

Miller (1993) explains that researchers in the field of psycho-biology have found that humans through their emotional experiences (like pain) get emotionally attached to objects. Therefore, societal institutions can be designed to channel these experiences towards mutually constructive rather than destructive behaviour. Such thinking could also support altruism over self-interested behaviour in society. Strangely this take on human nature could see altruism supporting Hobbes's (1968) famously narrow definition of human motivation.

While socio-biology provides original insights into the concepts of self-interest and altruism (see overviews in Wilson 2000, Hirshleifer 1977, Sesardic 1995 and Koslowski 1999) we find that behavioural economics, which examines the psychological aspects of economic behaviour, offers a similarly original and deep appreciation of these concepts. It is understood that psychological behaviour does have an impact upon the results provided by classical economics (Mullainathan and Thaler 2000, Kahneman and Tversky 1979, Smith 2005, Tomer 2007, Kahneman 2003a and 2003b). Interestingly, Adam Smith raises some behavioural-economics issues in both of his great books: for example, reference-dependence, loss aversion, inter-temporal choice and self-control, overconfidence and altruism (Ashraf et al. 2005). He explains loss aversion by saying that adverse incidents have a much greater impact on human psyche compared to positive incidents (Smith 1790, pp. 176-177). Recent research on both human behaviour (Chen et al. 2005) and from brain-imaging technology (O'Doherty et al. 2001) has shown the existence of loss aversion.

Chen et al. (2005, p. 517) undertake an experimental study with capuchin monkeys, finding that these monkeys react rationally and that their behaviour can be described by standard price theory. They find that even when there is a lack of social learning they found that when these monkeys showed human behavioural traits of loss aversion and reference dependence when they were faced with complex decisions. These findings show that these traits are a function of innate behaviour. O'Doherty et al. (2001) used functional magnetic resonance imaging (fMRI) technology to analyse the brain activation in the orbitofrontal cortex (OFC) in humans undertaking emotion-related tasks in which a correct choice led to a probabilistically determined financial reward and an incorrect choice led to financial loss. They found that specific parts of the OFC were activated in response to this activity. They also noticed that the extent of brain functioning in the OFC was related to the size of the reward/loss. This shows that there is a direct linkage between rewards and losses, human emotions and OFC brain functionality. Such brain activity could also relate to emotions that result from rewards/losses from self-interested and altruistic behaviour in humans.

There have been many advances in behavioural economics that have helped us to understand human behaviour better in economic and social situations. Some of these advances relate to an improved understanding of fairness and social preferences, reference-dependence and loss aversion, preferences over risky, uncertain outcomes and time-discounting with specific applications in macroeconomics, labour economics, finance and law (Camerer et al. 2004). Additionally, a deeper understanding of aspects of behaviour, such as bounded rationality (refers to the fact that human decision making and rationality are limited to the available information, cognitive intelligence and duration of time in order to make that decision), will-power and self-interest has helped us gain an improved understanding of how human behaviour relates to economic

outcomes Camerer (1999). Discussion in this section has shown that socio-biology and behavioural economics have a significant impact on the study of altruism and selfishness.

2.5 Neuroeconomic insights into altruism and selfishness

Neuroeconomics research that has been conducted recently challenges the classical economics concept of human behaviour and rationality (Neumärker 2007). We can find a useful outline of this new interdisciplinary area of economic and social behaviour, with some speculation on its future, in, for example, Rustichini (2005). Neuroeconomics has been explained by Glimcher et al. (2008, p. 7) as:

> Since the late 1990s a group of interdisciplinary scholars have begun to combine the social and natural scientific approaches to the study of choice into an emerging synthetic discipline now called Neuroeconomics. The central assumption of this discipline is that by combining both theoretical and empirical tools from neuroscience, psychology and economics into a single approach, the resulting synthesis will provide insights valuable to all three parent disciplines. Studies conducted to date seem to support that conclusion. Theories from economics and psychology have already begun to restructure our neurobiological understanding of decision making, and a number of recent neurobiological findings are beginning to suggest constraints on theoretical models of choice developed in both economic and psychological domains.

Fehr and Camerer (2007) explain how neuroscience and economics can provide a powerful explanation of human social interaction (see also Fehr and Rockenbach 2004, Simon 1992, Glimcher et al. 2008, Levine 2006, Mayr et al. 2008, O'Doherty et al.

2001, Singer and Fehr 2005, Colin 2007, Yu and Zhou 2007, Zak 2007). They state that social preference theories can explain brain activation related to altruism, fairness and trust that are associated with cognitive control, processing of emotions and the integration of costs and benefits, which is consistent with the resolution of conflicts between self-interest and other motives. Levine (2006) and Lynne (2006) state that humans are not only self-interested, but they also care about the group of people around them. As Cory (2006, p. 592) neatly explains that this has 'profound implications':

> New findings in brain physiology, especially evolutionary neuroscience ... show that the transactional commercial market evolved from the interplay of our self-preservational (egoistic) and affectional (empathetic) neural circuitries. These fundamental brain circuitries, under homeostatic physiological regulation, are the neural substrate of our human social exchange activity – from sharing in primitive families to the gift exchange economy to the commercial market. Current microeconomic theory is structured on the assumption of a sole primary self-interest motive. The presence of the dual physiological motives, however, is clearly demonstrated in demand, supply, and equilibrium curves as well as in the basic calculus of price theory. This confirmed duality of motives opens the way for new and productive directions in research.

Similarly, work by Sanfey (2007), Singer and Fehr (2005), Zak (2007), Yu and Zhou (2007) and Neumärker (2007) provide us with other models explaining competitive bargaining behaviour and help us understand empathy and trust further.

Zak et al. (2005) have found that humans show significant trust with strangers. This occurs due to the oxytocin hormone that facilitates social recognition and trust in humans. Krueger et al. (2005) also show that the tegmental and septal parts of the brain can support conditional and unconditional trust between human beings. Krajbich et al.

(2009) on the other hand, have found that damage to the ventromedial prefrontal cortex in a human brain causes dysfunctional real-life social behaviour. They explain that lower levels of trustworthiness and lower insensitivity to guilt felt by humans with such conditions. Such work in neuroeconomics show how Adam Smith's self-interest and altruism traits are affected by human brain physiology and neuroscience.

2.6 Defining selfishness and altruism

It is important to define the concepts of selfishness and altruism in this chapter. As, Eldakar and Wilson (2008, p. 6985) state that it is rather hard to obtain exact definitions of the terms selfishness and altruism. However, the *Dictionary of Psychology* defines the term selfishness as 'exaggerated regard for personal advantage, accompanied by a disregard for the welfare or happiness of others' (Corsini 2002, p. 878).

On the other hand, Gintis (2010, p. 9) provides a definition for altruism which states that, 'The behaviour of an individual is altruistic if it benefits other members of the group and the individual would increase his own payoff by switching to another behaviour.' The Nobel economic laureate Herbet Simon (1992, p. 73) agrees with Gintis' (2009) definition on altruism and in the context of evolutionary theory he states that 'Altruism is defined in evolutionary theory as behaviour that sacrifices one's own production of progeny, one's own fitness, to enhance the fitness of others.' Bowles (2008, p. 326) also agrees with the definition of altruism provided by Gintis (2009) and Simon (1992) when he says that, 'Altruism is conferring benefits on others at a cost to oneself.'

In essence, based on the definition above, selfishness occurs when a person puts his or her interest ahead of the interests of others and altruism can be defined as the cost

that a person will pay to provide a benefit to another person. These definitions of selfishness and altruism are broadly in line with Adam Smith's concepts of self-interest and beneficence in his two seminal books (Smith 1776 and 1790).

2.7 Conclusion

This chapter began with a discussion on Adam Smith's view on self-interest and altruism based on his books, *An Inquiry into the Nature and Causes of the Wealth of Nations* (Smith 1776) and *The Theory of Moral Sentiments* (Smith 1790) respectively. Morality has been the centre point of Adam Smith's teachings and this chapter shows how self-interest and altruism interact within the field of moral philosophy, leading to a discussion on the contributions of socio-biology, behavioural economics and neuroeconomics to a discussion on Adam Smith's ideas of self-interest and altruism.

This chapter has provided a review of Adam Smith's views of self-interest and altruism. The follow chapter will extend this discussion by outlining the latest research on a concept called strong reciprocity. Strong reciprocity is a concept coined by Bowles and Gintis (2004), who refer to second-order altruistic behaviour in which an individual punishes selfish individuals who do not follow the group's social norms at a cost to themselves. In essence, the next chapter will integrate the concepts of self-interest, altruism and punishment.

CHAPTER THREE

Understanding strong reciprocity

3.1 Introduction

The previous chapter reviewed the literature on Adam Smith's concepts of self-interest and altruism. In this chapter, these concepts will be developed further in relation to an additional concept known as 'strong reciprocity'. The latter term has been defined as second-order altruism, whereby an individual, at a cost to themselves, punishes selfish individuals who do not follow the group's social norms (Bowles and Gintis 2004). Eldakar (2007) builds on the concept of strong reciprocity by introducing the concept of 'selfish punishers', who manifest a form of second-order altruism by punishing other selfish people to increase their own pay-offs in future social interactions, in contrast to strong reciprocators. After discussing the concepts of strong reciprocity and selfish punishment this chapter will also briefly review the insights that behavioural economics and neuroeconomics can provide to strong reciprocity.

3.2 What is strong reciprocity?

As discussed in the previous chapter, and as argued by, for example, Ashraf et al. (2005), Montes (2003) and Evensky (2005), Adam Smith understood selfishness and altruism as being complementary in nature. According to Smith (1776, 1790) every individual possesses the behavioural traits of selfishness and altruism to some degree. However, it is possible that an individual can become overly selfish to a socially unacceptable extent. Smith supported the concept of punishment as a way of restricting

such high levels of selfishness. This conception of punishment is the basis for an understanding of strong reciprocity.

Strong reciprocity is a concept coined by Gintis (2000, p. 178) who defines it as 'Strong reciprocity is a form of altruism, in that it benefits group members at a cost to the [altruistic punishers] strong reciprocators themselves.' Bowles and Gintis (2003, p. 429) refine this definition by stating:

> We hypothesize that where members of a group benefit from mutual adherence to a norm, individuals may obey the norm and punish its violators, even when this behavior incurs fitness costs by comparison to other group members who either do not obey the norm or do not punish norm violators, or both. We call this strong reciprocity. Strong reciprocity is altruistic, conferring group benefits by promoting cooperation, while imposing upon the reciprocator the cost of punishing shirkers.

This definition of strong reciprocity is reiterated by Bowles, Fehr and Gintis (2003, p. 1) who identify the concept thus:

> Strong reciprocity is a combination of altruistic rewarding, which is a predisposition to reward others for cooperative, norm-abiding, behaviors, and altruistic punishment, which is a propensity to sanction others for norm violations. Strong reciprocators bear the cost of rewarding or punishing but gain no individual economic net benefit from their acts.

Additionally, Bowles and Gintis (2004) explain strong reciprocity as second-order altruistic behaviour, whereby an individual punishes selfish individuals who do not follow the group's social norms, at a cost to themselves. They also find that strong reciprocity exists even in groups that are unrelated and that this behaviour cannot be

explained by theories of kin selection, reciprocal altruism, costly signalling or indirect reciprocity.

In contrast, Eldakar and Wilson (2008, p. 6982) differentiate 'selfish punishers' from the concept of strong reciprocity and state that:

> Recent models have explored punishment as an important mechanism favoring the evolution of altruism, but punishment can be costly to the punisher, making it a form of second-order altruism. This model identifies a strategy called 'selfish punisher' that involves behaving selfishly in first-order interactions and altruistically in second-order interactions by punishing other selfish individuals. Selfish punishers cause selfishness to be a self-limiting strategy, enabling altruists to coexist in a stable equilibrium. This polymorphism can be regarded as a division of labor, or mutualism, in which the benefits obtained by first-order selfishness help to 'pay' for second-order altruism.

Eldakar et al. (2007) suggest that strong reciprocity is a second-order altruism, whereby the selfish individual punishes other selfish people, in order to increase his or her own future payoffs. Eldakar and Wilson (2008) and Eldakar et al. (2007) differentiate their concept of selfish punishment from strong reciprocity by stating that in this case selfish individuals punish other selfish individuals to increase their pay-offs in future social interactions, as compared to Bowles and Gintis (2003), who discuss altruistic individuals punishing selfish individuals to increase altruism within the group.

When considering strong reciprocity, we recall Adam Smith's impartial spectator (as discussed in *The Theory of Moral Sentiments*, 1790). Smith believes that this impartial spectator helps us view problems and make decisions from the perspective of a third person (i.e. a bystander) who is impartial to the situation at hand. The conception of punishment meted out by a strong reciprocator aligns with Adam Smith's conception

of the impartial spectator, who views selfishness as an impropriety (Smith 1790, p. 16). An 'impartial spectator' effectively acts as an umpire for a person's own actions. However, the literature has also stated that strong reciprocators are willing to punish others at their own cost. This is supported by a quote from Bowles and Gintis (2003, p. 429):

> We hypothesize that where members of a group benefit from mutual adherence to a norm, individuals may obey the norm and punish its violators, even when this behavior incurs fitness costs by comparison to other group members who either do not obey the norm or do not punish norm violators, or both. We call this strong reciprocity. Strong reciprocity is altruistic, conferring group benefits by promoting cooperation, while imposing upon the reciprocator the cost of punishing shirkers.

3.3 How does strong reciprocity relate to selfishness and altruism?

Altruism and selfishness are directly linked to the concept of strong reciprocity. It is important to explicate both altruism and selfishness further before we can discuss how they explain strong reciprocity.

According to Smith (1790, p. 2) altruism is a natural trait in human beings. He argues that humans can empathise with other people's sorrow and that every human has the capability to be at least a little altruistic. Further, it has been argued that altruism maximises benefits from the perspective of a group or society. Andreoni (1995) shows that co-operation in games does not occur due to error or confusion between free riders (where free riders can be defined as selfish individuals who make use of altruistic individuals for their own benefit without regard for the altruistic individual). He provides examples of donations to the Red Cross and other charitable organisations that

occur due to this conscious action of kindness or altruism towards our fellow human beings.

Andreoni (1995) shows that both conscious altruism and errors in judgment contribute to altruistic behaviour within a game and that people who are consciously altruistic know that they can free ride but opt not to do so due to their altruistic nature. As learning takes place between individuals, the level of altruism increases very quickly and the reduction in the level of errors leads to a co-operative equilibrium resulting from greater altruism and lower errors in judgment.

In addition to direct reciprocity as explained by Andreoni (1995), it can be seen that indirect reciprocity also occurs through image scoring, that is when people are altruistic to other people because of the altruistic actions taken by the first person. For example, if a person has helped another person, then there is likelihood that this second person will also help the first person in return (Wedekind and Milinski 2000). By contrast, Andreoni (1988) believes that a pure public goods approach to altruism is limited for four reasons. Firstly, because free riding is dominant in altruistic societies, resulting in the level of altruism decreasing to close to zero and a situation whereby only the richest people tend to contribute to the public good. Secondly, because a pure altruism model showed strong neutrality results, whereas the Nash equilibrium was independent of the distribution of income, direct government provision and distorting subsidies. Thirdly, because exogenous increases in altruism do not change the level of perception regarding total equilibrium donations. Lastly, because the change in public goods is invariant to redistribution, joint provision and changes in population. He believes that a new approach should be taken, but he questions whether a co-operative equilibrium can be sustained in a large economy.

Sugden (1984), Margolis (1984), Bernheim et al. (1985), Andreoni (1993) and Olson (1971) on the other hand propose that emotions such as guilt, repentance, envy, sympathy, emulation or a taste for fairness may provide selective incentives for giving. Andreoni (1993, p. 1317) contributes to this discussion through the example of how taxation causes 'incomplete crowding out' for public goods because it creates a lower bound for altruistic giving, creating a situation where people contribute a larger amount voluntarily. Also, endogenous giving between people increases altruism (causing a crowding in effect) more than exogenous giving by the government (reducing the crowding out effect) which increases altruism but to a lesser extent than reducing taxation. Andreoni (1993) further identifies gaps in our knowledge of the effects of social pressure, peer group effects and theories of fund raising, believing that such processes could manipulate or change people's preferences for giving.

We have analysed altruism and implicitly its effect on the problem of social cost, and now have to discuss the other side of the equation, relating to selfishness. Selfish (or cheating) strategies in a game maximise an individual's payoff. While it helps individuals to be selfish, it is important to note that this trait also exists side by side with altruism in society (Ashraf et al. 2005). However, Fehr et al. (2002) and Eldakar et al. (2007) believe that it is difficult to increase the level of altruism in the presence of selfish cheaters. The situation changes when members of a society are able to punish cheaters (usually in the form of excluding cheaters from sharing the group's common resource pool). However, finding and punishing cheaters comes at a cost. Here, Nakamura and Iwasa (2006), Sigmund et al. (2001) and Eldakar et al. (2007) identify a negative relationship between punishment and altruism.

Fehr and Gächter (2000, p. 980) state that 'free riding causes negative emotions' that could result in punishment being applied even if this is costly. Such punishment

causes the level of selfishness to reduce and that of co-operation to increase in the group. In this context the greater the degree of free riding the more likely it is that the free riders are punished. This argument is also supported by laboratory experiments undertaken by Price et al. (2002). Similarly, a laboratory experiment conducted by Burnham and Hare (2005) that used a robot to monitor every individual's actions within one of two groups of participants found that greater co-operation developed in the group that was monitored compared to the group that was not monitored by the robot. McCabe et al. (1996) also demonstrated that reciprocity exists in repetitive games and found that some people are prone to co-operative behaviour, as they co-operate in single play games as if they are in a repetitive series of different games. Mendes (2004) explains that in small groups that have collective monitoring, levels of strong reciprocity can differ depending on the intra- and intergroup dynamics between the members of the group. In a larger society where collective monitoring cannot successfully prevail, clustering (where individuals collectively support the idea of punishment) can be a more successful factor that can maintain strong reciprocity compared to collective monitoring.

Regardless of the fact that punishment can result in ostracism of selfish individuals, Hauert et al. (2002) at first acknowledged that defection is the dominant strategy. However, using a replicator dynamic model, they found that if co-operation can be developed within an iterative game then a cyclical equilibrium between defection and co-operation exists, leading to a sizeable average level of co-operation within the population. Here, McCabe, Rigdon et al. (2003) add that population clustering in Prisoners' Dilemma games can allow for some co-operative strategies to develop within populations of stable defecting strategies, resulting in higher levels of co-operation being sustained in public goods games (where public goods can be defined as goods

provided for societal use without any intent to profit, for example, public libraries and public goods games relate to the fair distribution/use of public goods between different people). The findings of McCabe et al. (1996) also support the argument that reciprocity in repetitive games increases co-operation. Bowles and Gintis (2002, 2004) have argued that it is more than self-interest that requires humans to punish free riders. They contend that humans will by contrast altruistically punish free riders in order to increase co-operation within the group, even if they do not directly benefit from such behaviour.

Strong reciprocity can support the punishment of free riders effectively and maintain a high level of co-operation if the frequency of reciprocators is not too low or the group is not too large. Strong reciprocity can be effective in making free-riders contribute towards the public good, provided that it does not crowd out the pre-existing social preferences that assisted altruistic behaviour prior to the application of punishment (Bowles 2008, Bowles and Hwang 2008). When considering a team environment, each individual will share the monitoring and punishment cost, reducing the overall burden on altruistic punishers.

3.4 How do genetic co-evolution, co-operation and other factors affect strong reciprocity?

Bowles and Gintis (2002b) explain that the existence of groups has been an important factor in human evolution. Multi-level selection and the dynamics of gene culture co-evolution have also been highly significant in developing co-operation between human beings. Bowles and Gintis (2002b) also believe that relatedness, repeated games and other aspects of social interaction among group members might provide higher fitness

to those that show unselfish behaviour. However, these researchers also note that individuals with higher fitness will most likely survive. As DS Wilson and EO Wilson (2007, p.328) point out, 'Selfish individuals might out-compete altruists within groups, but internally altruistic groups out-compete selfish groups. This is the essential logic of what has become known as multilevel selection theory.' They also state that:

> The whole point of multilevel selection theory is, however, to examine the component vectors of evolutionary change, based on the targets of selection at each biological level and, in particular, to ask whether genes can evolve on the strength of between-group selection, despite a selective disadvantage within groups. Multilevel selection models calculate the average effects of genes, just like any other population genetics model, but the final vector includes both levels of selection and, by itself, cannot possibly be used as an argument against group selection. (Wilson and Wilson 2007, pp. 335-6)

Dawkins (1976) believes that people are born with the trait of selfishness, as selfishness is an important enhancer of a gene's success and this process instils selfishness in individual behaviour. Therefore, he believes that we are born selfish, but we can learn generosity and altruism. Bowles and Gintis (2002b) question Charles Darwin's thoughts regarding selfishness. Darwin (1997) explains that individuals in a group can caution others of impending danger in order to protect themselves and the group from harm. Such actions are identified as selfish behaviour according to Bowles and Gintis (2002b) when they consider that 'tribes in which these behaviours were common would spread and be victorious over other tribes'. Similar to the principle of self-interest, they propose that natural environment and gene interactions affect the evolution of cultures, while culture affects natural and social environments in which relative fitness of genetically transmitted behaviour traits are determined. Compelling

empirical evidence provided by researchers including Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1988) and Durham (1992) suggests that culture affects genetic evolution.

Bowles and Gintis (2002b) explain that this interaction occurs through the evolution of genetically transmitted behaviours that relate to individual growth along with culturally transmitted behaviours that are passed on through group level activities and social norms. In contrast some models explain how cultural factors such as resource sharing are critical to the evolution of genetically transmitted altruism through natural selection. Odling-Smee, Laland and Feldman (2003) and Bowles (2000) believe that it may be helpful to represent human cultures and institutional structures as supporting the 'creation of a particular environment'(Bowles 2000, p. 148), which in turn affects genetic evolution.

It is important to question how high levels of co-operation are maintained despite low levels of genetic relatedness. Bowles and Gintis (2004) explain that humans can punish others at a personal cost. This type of altruistic punishment has been observed to maintain high levels of co-operation in experiments. Bowles and Gintis (2000) provide a model of co-operation and punishment, describing this altruistic punishment as strong reciprocity. Here, group members adhere to social norms and strong reciprocators punish violators, even when these punishers receive lower pay-offs than other group members (who are selfish or altruistic but do not punish) due to the cost of punishment.

Fehr and Fischbacher (2002, 2003, and 2005) explain that both altruism and selfishness are fundamental to 'our evolutionary origins [and] ... social relations' (2003, p. 785). According to them, gene based evolutionary theories cannot explain interactions between altruists and selfish individuals. Instead, a combination of cultural evolution and gene-culture co-evolution theories are required to understand this issue

further. Laboratory experiments (Fehr and Gatcher 2002, Ostrom et al. 1994) and field data (Boehm et al. 1993) have shown that individuals punish non co-operative behaviour even in one shot interactions. Even though such altruistic punishment increases co-operation in a group it creates a dilemma, as existing models suggest that altruistic co-operation between non-genetically related individuals is evolutionarily stable only in small groups. Therefore, the effects of punishment in one shot experiments leads to predictions that people will not suffer costs to punish other individuals to provide a benefit to a larger group of non-genetically related individuals. Relating Ostrom et al. (1994) back to Cooper and Wallace (2004), Ostrom et al. (1994) state that small groups are successful in increasing altruism (where group members have been together during the game), while Cooper and Wallace (2004) review a different condition – where group members have been secluded for multiple periods. In this case, the size of the group, cost of being altruistic and the number of periods are important factors. They find that medium size groups do better than smaller groups.

Boyd et al. (2003) demonstrate that an asymmetric relationship between altruistic co-operation and punishment allows altruistic punishment to evolve in larger populations even in one-off interactions. This allows for altruistic co-operation and punishment to exist simultaneously even in large groups, under other parametric values approximate conditions that can exemplify cultural evolution in small-scale societies. In comparison, Masclet and Villeval (2006) analyse the rationale of costly punishment in non-genetically related groups, investigating inequality aversion and negative emotions between individuals as possible determinants for punishment. Their results indicate that the intensity of punishment increases with the level of inequality and reduce the level of fitness between the individuals over subsequent games.

Human co-operation seems to be an evolutionary puzzle, as people voluntarily participate in expensive co-operation by punishing non co-operators. Even when, as Burnham and Johnson (2005) state, this cost cannot be recovered by kin selection, reciprocal altruism, indirect reciprocity or costly signalling. They believe that strong reciprocity is not a newly documented concept. It is maladaptive and has evolved by individual selection. Group selection in their view could possibly play a role, but it is neither necessary nor sufficient to explain co-operative behaviour in humans. In addition, contemporary behavioural theory relating to strong reciprocity has been developed by Cosmides and Tooby (1992), Dawkins (1976), Maynard Smith (1982), Williams (1966) and Wilson (2000), who argue that non-kin relations can be modelled using self-interested actors.

Since the publication of these works substantial behavioural theory research has been undertaken to understand strong reciprocity resulting from non kinship, family and sexual relations. Colman (2006) provides a brief background to co-operation and strong reciprocity. Giving an example of a bird that emits an alarm when spotting a predator, he believes that according to Dawkins (1976) this bird would not survive compared to the bird with the 'selfish gene' that saves its energy and avoids being spotted by the predator by not emitting an alarm. Fehr and Gächter (2000) support Colman's (2006) intuition by showing that co-operation can exist between non-related individuals. Nonetheless, Colman (2006) believes that while strong reciprocity is the best explanation for co-operative behaviour between humans, he also believes that due to the cost of punishment leading to lower fitness strong reciprocators would not survive over time and will eventually be eliminated.

In contrast to Coleman's argument, in biological and psychological theories of evolution of co-operation and reciprocity cheater detection is a crucial task. Cheater

detection is important in a social co-ordination function within a group, whereby dominance provides priority access to resources and provides a set of social norms that need to be followed within the group. As a result, individuals on top of an organisation hierarchy can observe and punish individuals lower down the hierarchy as they detect cheating better than them rather than vice versa, according to dominance theory (Cummins 1999).

Thus, positions higher up in a hierarchy are attained by setting up coalitions with non-genetically related individuals. Social exchange theory (i.e. a theory of co-operative effort for mutual benefit) on the other hand explains cheater detection in such mutually beneficial relationships (Cummins 1999). For instance, in chimpanzees (like in other species of social animals and non human primates) it is seen that cheating can cause a termination of alliances (de Waal 2000), although the amount of cheating is balanced out by the social rank of the individual (Chapais 1992, Cheney 1983) as higher ranking individuals will usually get away with more cheating due to their social rank.

In contrast, where dominance or social exchange theories are not relevant and when both individuals have equal standing in the game and a partial equilibrium exists in a two-person game, both individuals will likely have an asymmetric view of their position in this game. These individuals will negotiate and improve their pay-offs if they do not have similar views. Here, co-operation seems to emerge usually through unilateral concessions. If each individual's views are diverse, then there is a strong likelihood of a conflict developing that will reduce the co-operation within the game as well as the payoff of each individual (Caruso 2008).

There has been substantial research in understanding social behaviour through experimental, theoretical and empirical research in strong reciprocity. However, Gintis et al. (2003) believe that substantially more experimental and theoretical work needs to

be done to understand human pro-social behaviour. They propose a course for this future work by arguing that further insight needs to be obtained with newer models that improve on the existing economic models of self-interested actors and the biological model of self-regarding reciprocal altruists. There are numerous other factors that also affect strong reciprocity and additional research needs to be conducted to understand these factors and the relationships between them. Some of these factors are discussed below.

Learning is one of these factors that can affect human behaviour, and humans interacting within games show strong reciprocity in order to increase altruism as they learn about selfish people's actions in previous games. Calderon and Zarama (2006) analysed an ultimatum game and found that while the results were initially explainable by non co-operative game theory, when they introduced learning for each player, the outcomes started coming close to human behaviour. In this case, learning is a significant factor that could impact results in repetitive games as humans learn and change their behaviour, especially as they punish selfish individuals for acting in self-interested ways in previous games.

Trust is another factor that can impact on strong reciprocity, and this can be seen in an experimental trust game undertaken by Buchan et al. (2002), that examined the effects of country, social distance and communication on trust and reciprocity in China, Korea, Japan and the USA. They found, as expected, a negative relationship between trust and social distance and a direct correlation between trust and strong reciprocity. Also, personal rather than impersonal communication impacted on trust and reciprocity between players. They noticed that social distance could be another factor that affected strong reciprocity.

In response to work on strong reciprocity, Carpenter and Matthews (2004) have discussed the concept of social reciprocity. This concept differs from strong reciprocity in that social reciprocators will punish violators at some personal cost unconditionally, that is, without any dependence on future payoffs, revenge or altruism, when any social norm is broken (for example in the case of an individual free riding). Thus, social reciprocity is seen as a triggered normative response. Carpenter and Matthews' (2004) results show that reciprocators in a social setting do punish even if they face a cost.

Similarly, in an experiment undertaken by Carter and Castillo (2002), analysing the social capital aspects of altruism, trust and reciprocity in a 'sample of South African communities' (2002, p. 1), it was discovered that trust, reciprocity and altruism are related. However, it was noted that altruism is also clearly different to trust and reciprocity, though a relatively strong correlation between trust and reciprocity seems to exist. This indicated that when reciprocity norms existed then trust would be high within the community. Statistical data from the same communities showed that these norms have significant impact on household wellbeing. They had a positive impact on the wellbeing of urban communities and a negative impact on traditional rural communities.

While we have analysed the different factors that affect strong reciprocity, we also need to understand that these relationships can hold in different experimental settings. Dohmen et al. (2008) have analysed reciprocity in relation to 'labour market behaviour [and] ... overall satisfaction with life's outcomes' (2008, p. 10). They found that positive reciprocity existed between higher wages and working harder. They found negative reciprocity exists between reduced effort and lower pay. However, firms do not lower pay due to reduced effort, rather reduced effort results in unemployment because employers do not want to hire underperforming employees. They also found

that positive reciprocity exists between close friendships and greater happiness for people with life in general, and that people who did not have close friendships would be less happy due to the social nature of humans.

A similar study was undertaken by Fehr and Falk (1999), in which it was argued that workers underbid the prevailing wage rate 'under conditions of incomplete labour contracts' (p. 109), in order to gain employment when there is higher unemployment in the economy. Nonetheless, employers do not cut the wage rate below market rates primarily as work effort directly equates to worker's compensation and lower wage rates would mean that the worker has discretion to deliver lower work effort. In this case, the punishment occurs when the worker reduces his work effort with a decrease in wage rate. Similarly, employers punish their workers by cancelling their employment contracts if the worker does not perform to the expectations of the employer for the wage rate this worker is obtaining from the employer.

3.5 Is strong reciprocity affected by fairness, intentions and social preferences?

Thus far we have discussed how various factors can impact on strong reciprocity. Camerer and Thaler (1995) have explained that fairness is another factor that affects people's behaviour in relation to strong reciprocity. Based on empirical research on fairness perceptions of consumers, Kahneman et al. (1986) have observed that consumer behaviour is not shaped by pure selfishness related factors. For example, in a gift exchange game a larger gift provided by one person is reimbursed with a large gift from the person receiving the gift (Berg et al. 1995, Gächter and Falk 2001, Fehr et al. 1996).

Strong reciprocity can also have a strong effect on market forces within auction markets and it can determine the size of the gift that the first player will receive (Fehr and Falk 1999, Fehr et al. 1993, 1998), which means that if individuals are unfair then other people react adversely to their unfair act. Fehr et al. (1997) have also experimentally shown that the 'set of enforceable contracts' (p. 833) increases considerably with non-selfish behaviour. This is the case because although contracts are usually incomplete in nature, a combination of unselfish behaviour and strong reciprocity covers for the incompleteness and results in an increase in the number of such contracts. Akerlof (1982) and Akerlof and Yellen (1988, 1990) moreover believe that fairness is the possible explanation for the fact that wages are above market clearing prices, which supports Bewley's (1999) explanation that fairness causes wages to be resistant to drop during recessions, as reducing wages will reduce worker moral and overall productivity. Similarly, Rossi and Warglien (2000) analyse fairness in an agency theory framework with an N-person Prisoners' Dilemma game. They find that if principals (owners) are unfair, then agents (workers) defect and become less co-operative as well. Fairness consideration of the principal's actions can affect the behaviour of the agent and result in their propensity to either co-operate or defect in the game. This shows that fairness is an important factor in understanding strong reciprocity.

Bolton and Ockenfels (2000) also consider fairness (in the form of equity) and reciprocity, proposing a theory of equity, reciprocity and competition. Their model uses an N-person game where players are randomly drawn from the population. In their model, payoffs are non negative and players do not play with an opponent in more than one instance. According to them, fairness and reciprocity are primarily important in groups, and they believe that evolutionary biology research supports their finding that

individual success depends on group success, and so therefore groups have a tendency to punish free riders. Bolton and Ockenfels (2000, p. 189) Guth (1995), Ellingsen (1997), Kockesen et al. (2000) also support these conclusions based on evolutionary biology models.

We have seen that fairness is important when considering strong reciprocity, but Rabin (1993) explains that intentions are also an important factor to consider when people are motivated by reciprocal factors. He assumes that players in a two-person game experience psychological payoffs in addition to material payoffs, where these psychological payoffs depend on kindness (altruism), which in turn depends on beliefs. Geanakoplos et al. (1989) provide a framework of psychological game theory that incorporates intentions and reciprocal behaviour that cannot be shown in standard game theory models. Rabin (1993, p. 1296) adopts Geanakoplos et al.'s (1989) model but points out that it does not take into account dynamic or sequential games that are essential for applied research.

Prior to Rabin's (1993) model, traditional game theory could not provide accurate solutions to games that incorporate psychological behaviour such as intentions. In the same vein, Eckel and Wilson (1998) performed a game theory experiment in which they included 'reputational priors' (p. 1) that can be explained as the reputation gained by a person in previous games. In this game, each player was shown 'facial schematics [that embodied psychological] ... affective content' (p22). Introduction of these facial expressions changed the result of the game in ways that could not be predicted by mathematical game theory. Eckel and Wilson (1998) believe that psychological game theory would have to be used to solve such a problem. They do not know if the changes in game behaviour due to changes in facial expressions were due to stereotyping,

attribution or evolutionary psychology. Finding evidence of which of these might be a casual factor would add insight to psychological game theory.

Falk and Fischbacher (2006) develop this concept of intentions further by providing a formal theory of reciprocity that explains that in addition to actions, intentions are also important in judging a person's act of kindness. They believe that this theory is in line with the different types of experimental games such as the ultimatum game, the Prisoners' Dilemma and public goods games. This theory also explains why outcomes may seem to be fair in bilateral interactions, while they may seem unfair in competitive markets.

Social preference is another factor that can affect strong reciprocity. In analysing social preferences, Fehr and Fischbacher (2002) state that humans do not only care about material outcomes, but also about the positive or negative social preferences that are associated with these outcomes. Economics fails to consider these social preferences, legal aspects of competition and collective action, determining factors related to material incentives and issues relating to optimality of contracts and property rights, including factors that affect social norms and market forces within an economy.

Social norms define the behaviour that are based on shared beliefs within a group and require individuals to follow that behaviour (Voss 2001). Social norms have been an unsolved problem in cognitive social science directly affecting game theoretic research. Sanctions, an important mechanism to support norm enforcement, are 'largely driven by non-selfish motives' (Fehr and Fischbacher 2004a, p. 185). Despite some recent progress in the analysis of social norms (Henrich et al. 2001, Hechter and Opps 2001, Posner 2002) there is a lot more research required in understanding the effect of emotions on co-operative behaviour and punishment decision making as well as the neural reinforcement of social norms (Rilling et al. 2002, Fehr and Gächter 2002,

Sanfey et al. 2003). At present, even though the socio-economic environment seems to define the costs and benefits related to co-operation and punishment, there still seems to be a lack of knowledge in relation to the social and economic determinants of social norms (Fehr and Fischbacher 2004a).

Leading on from this discussion, Fehr and Fischbacher (2004b) make the conjecture that third parties who are unaffected by the violation of social norms are sometimes willing to enforce punishment on norm violators at a personal cost. They find that majority of the parties punished norm violation with the punishment increasing in line with the norm violation. Their results show that strong reciprocity extends to the sanctioning of non co-operative behaviour even with unaffected third parties, while second parties whose economic payoffs are directly affected understandably punish norm violators more strongly than unaffected third parties.

We have discussed strong reciprocity as a second-order altruistic trait and subsequently analysing the factors that affect strong reciprocity. It was pointed out that Bowles and Gintis (2000) considered strong reciprocity to be second-order altruism. Eldakar (2007) has since shown that it is actually second-order altruism, as these strong reciprocators are eliminating other selfish individuals to reduce the number of selfish individuals in subsequent games and resultantly to increase their own payoffs in those games. Eldakar et al.'s (2007) model is discussed in more detail below to understand the concept of second-order altruism.

Eldakar et al. (2007) develop a two-phase evolutionary model of an 'infinite population of individuals' (p. 199) with varying propensity for altruism and punishment. In phase one each individual allocates a portion of their private empowerment, based on their level of altruism, to the common pool. At the end of phase one, this common pool is doubled and equally redistributed to all the individuals

who donated funds to this common pool. Selfish individuals allocate the minimum level of their personal wealth into the common pool, which is just sufficient enough to make them eligible to receive an allocation from the common pool at the end of phase, thereby maximising their personal wealth.

During phase two, individuals (regardless of their level of altruism) can allocate resources to find and punish cheaters. The level of resources that are allocated by individuals in the game is proportional to, firstly, the individual's punishment trait, secondly to the average amount of cheating that took place within the group and thirdly to the amount required to detect the worst cheater. In the event that the most selfish individual is detected they are excluded from future rounds of the game. This cheater will then be replaced by a new individual drawn randomly from the same pool of resources from which the initial individuals were selected.

Eldakar et al. (2007) has set up a game with a large number of randomly formed groups that is played numerous times. Individual fitness is evaluated based on total earnings and baseline fitness value (showing that fitness cannot be assessed primarily on interactions within the game). Fitness is the combination of abundance and fitness for each strategy type, providing a cumulative fitness value for the 121 types identified in this study. The fitness value is then normalised to equal the value of '1' that represents the frequency of each type of strategy. Therefore, each individual has an updated fitness at the beginning of each round in the game.

Further, the most selfish individual (i.e. the worst cheater) will attempt to identify the second worst cheater, so that the most selfish individual does not get excluded. This happens as it reduces the amount of cheating perceived by the group resultantly reducing the strength of punishment and as a result decreases the chance that the worst

cheater is excluded in future rounds. Within this environment it is found that over time selfish individuals are eliminated from the game.

Eldakar et al. (2007) perform two simulations in this study to identify the relationship between altruism and punishment:

(i) They have an equal number of altruistic and selfish individuals comprising the game. In this case, they find that altruism increases reasonably quickly, as selfish punisher punish selfish non-punishers by excluding them from subsequent games.

(ii) The level of altruism and punishment within the group is set to zero (A=0 and P=0). In this case they need to have at least a certain level of mutation ($M=10^{-4}$) in order to meet the threshold where punishment can result in increasing altruism (the same happens if the cost of punishment is too high). Below, this threshold there is insufficient amount of punishment in the game – resultantly altruism will take substantially longer to build.

Eldakar et al. (2007) also identified three main factors that affect the speed at which altruism increases in a dynamic game being: firstly, the cost of punishment, secondly, group size and thirdly the number of repetitive games (round length). The results indicate that the higher the cost of punishment the more likely selfish punishment is required, at lower costs the correlation between altruism and punishment is zero. Similarly, they found that group size is important, as the 'levels of altruism and punishment decline[s]' (p. 202) as group size increases and then starts to increase after a group size of seven individuals is attained. The number of repetitive games was also found to be important, due to the fact that individuals cannot be punished in a one-off game. However, as the number of games increased there was found to be a higher number of selfish individuals eliminated through punishment. These findings show that

the number of games has an indirect impact on the level of altruism and punishment in games, as increasing the number of rounds in a game decreases the cost of punishing cheaters.

3.6 Reviewing strong reciprocity through neuroeconomics

This preceding section has analysed strong reciprocity from the point of view of behavioural economics. Neuroeconomics has recently started providing a significant insight into strong reciprocity as well. We have noted that Camerer (1997, 1999, 2003a and 2003b), Benoit (2007), and Camerer et al. (2004) have argued that behavioural economics can help improve the realism of the psychological assumptions underlying economics and that these psychological factors affect strong reciprocity. It is therefore important to understand how neuroeconomics affects strong reciprocity. Klein (2008) provides a broad overview of the area related to neuroeconomics, social norms and strong reciprocity.

Reviewing neuroeconomics and strong reciprocity research we find that de Quervain et al. (2004) previously undertook an experiment to understand the neural basis related to altruistic punishment. They found that the dorsal striatum (interior part of the forebrain that is responsible for decision making in humans) gets activated which process rewards that result from goal-directed actions. Implications for this study of the study by Quervain et al. (2004) are that humans who show a stronger activation of the dorsal striatum in anticipation of punishing defectors are more likely to punish defectors than those humans who do not show this biological change.

In addition to this work, Krueger et al. (2008) have analysed game theory from a neuroeconomics viewpoint and specifically reviewed the concept of trust in such

games. Rilling et al. (2008) on the other hand have reviewed social interactions and human decision making from a neurobiological point of view. They argued that while human social environments are highly complex, they believe that researchers have already started understanding neural correlates of human decision making in reciprocal exchange and bargaining games through the functional neuroimaging, transcranial magnetic stimulation, and pharmacological manipulations that have been done earlier. Hagen and Hammerstein (2006) have also explained that the economic and biological models need to be synthesised to incorporate human cognition.

3.7 Conclusion

The previous chapter discussed how Adam Smith's work is related to the concepts of selfishness and altruism. This chapter has explained how the concept of strong reciprocity has been developed from these aspects of Smith's work. Bowles and Gintis (2000) first developed this concept of strong reciprocity. Strong reciprocators are second-order altruists; they are altruists who punish selfish individuals at a cost to themselves which results in an increase in altruism within the game. Subsequently, numerous other researchers have developed this concept, including research that reviews strong reciprocity from a behavioural economics and neuroeconomics viewpoint. However, Eldakar et al. (2007) have built their model that develops the concept of strong reciprocity further, in that they have introduced the concept of selfish punishers. The latter involves second-order altruism, in which selfish individuals punish other selfish individuals in order to increase their own future pay-offs.

The next chapter will take us through a review of game theory and complex systems, which will be used as tool to develop a model presented in chapter five, that is used to generate the results of the present study.

CHAPTER FOUR

Game theory and complex systems

4.1 Introduction

This chapter outlines game theory and complex systems and relates these areas to the methodology of the thesis. After providing a background on game theory, the chapter then briefly discusses the inter-linkages between game theory, behavioural economics and social cost. It then moves on to review three related complex system areas, these being ecological-economic systems, agent-based computational economics and agent-based social simulations. Following this review, the chapter analyses how complex systems relate to the concept of strong reciprocity and discusses more specifically the relations between complex systems models in game theory and strong reciprocity. These discussions directly relate to the 2-Person Random Interaction Model (2PRIM) model presented in the following chapter. Axelrod's (1984) model is briefly discussed, as it was the first complex-system paper relating to two-person games. Additionally, research by Bowles and Gintis (2000), Boyd et al. (2003) and Eldakar et al. (2007) is discussed in depth to provide a basis for the discussion in chapter five, where the 2PRIM will be developed and its results analysed.

Game theory is a research area within the field of applied mathematics within which specific applications have been developed to solve economic problems. Over the past few decades researchers from biology, politics, ecology, anthropology, environmental science, sociology and other areas have also used it to solve problems in their respective disciplines. As game theory has been applied widely to different areas, researchers have also realised that game-theoretic solutions cannot explain human

behaviour accurately. As a result, over time there has been an application of psychology to game-theoretic frameworks.

The concept of complex systems has been used to increase the usability of game theory by helping to analyse more computationally extensive problems or to depict systems with higher complexity. This chapter also discusses specific strong reciprocity models that provide a basis for the discussion in the next chapter.

This chapter in essence then reviews the tools, that is, game theory and complex systems, that are used to analyse the theoretical strong-reciprocity problem modelled subsequently in this thesis.

4.2 Game theory

Game theory is a research area in applied mathematics that is dedicated to resolving economic problems in both economic and social contexts. Contemporary game theory has two forms: non co-operative game theory (Nash 1951) and co-operative game theory (Shapley 1953, 1977, Shapley and Shubik 1954, Luce and Raiffa 1957, Aumann and Drèze 1974, Myerson 1977). These two forms provide the initial basis for all research in the area, including bargaining and auction theory. It is also relevant to note that non co-operative game theory was developed out of a previous theory provided by Von Neumann and Morgenstern (1944) called zero-sum game theory.

The two methods to analyse games are the axiomatic and strategic form (Von Neumann and Morgenstern 1944, Nash 1950, 1951, 1953). An axiomatic approach (sometimes called co-operative theory) provides a set of beneficial axioms that imply a unique solution. The strategic approach focuses on the outcomes of players in a non co-operative game (or strategic interaction), modelling the game theoretic-process over

time. The axiomatic approach was dominant until 1980, but Rubenstein's (1982) solution of alternating players with discounting utility over time provided the impetus for the strategic approach to gain momentum. Some researchers, including Myerson (1979, 1984, 1985) have tried to combine both the axiomatic and strategic approaches. However, these game theoretic forms require complete information, while most game-theoretic situations may only have incomplete information.

A substantial amount of research has taken place in the area of game theory over the past 70 years. Harsanyi (1966, 1967, 1968a, 1968b) has provided a strategic game-theoretic solution with incomplete information, and Chaterjee and Samuelson (1983) have used Harsanyi's theory to develop a static-bargaining model with incomplete information and with unknown reservation prices for a buyer and seller in a market. Rubenstein (1985) built on their model by using the discount utility factor of an opponent in a strategic interaction to replace unknown information. Around this time, Selten (1975) put forward sub-game perfect equilibrium game theory to explain how a less-than-perfect game equilibrium can provide for an overall equilibrium in a game. Leading from Selten's (1975) research, Kreps and Wilson (1982) have derived, utilising statistical inference, a model showing sequential equilibrium with incomplete information.

Game theory has developed substantially over the past few decades. This includes critical ideas such as those of Sen (1977), who has raised the question of how anyone could compare utility payoffs in games and state that a better or worse payoff has been received by one individual compared with another. The long-rehearsed debate in welfare economics on 'utility' relates directly to a discussion on individual preferences. There is still disagreement among economists regarding the concept of preferences and how they should be interpreted (see for example, Sen 2009, Nussbaum 1997).

Technically, subject utility cannot be compared between individuals, though some interpersonal comparisons can reasonably be made (Doughney 2002). Robinson (1962, p. 66) has explained that, with regard to the concept of utility as being 'locked in the individual's subjective consciousness', that it is not a unit at all. This would mean that it would be hard to compare utility across individuals. Sen (1977) adds to this discussion by arguing that preferences are complex and that people do not always make decisions that increase their own welfare or utility. Individuals may act against their personal welfare if they have a commitment that requires them to overlook their own interests (see also Searle 1990). The question then becomes: if different people have incommensurable preferences, how can we decide what is really the best payoff? If such a situation exists, then it would be rather hard to understand the payoffs in game-theoretic scenarios.

The critical arguments above are strong. However, in order to undertake the modelling exercise below, this thesis will use the standard definition of utility that considers utility to be a 'unit of value' that is comparable across individuals. Accepting the game theorist's view of utility means that preferences are taken to be ordinally ranked. I use, for convenience, the standard game theory definition of utility (i.e. ordinal preference ranking) in order to be able to have a quantitative way to compare individual satisfaction (utility) as well as the levels of altruism and selfishness in a game-theoretic scenario. In effect, assuming both commensurability and an ability to make interpersonal comparisons of utility (welfare, satisfaction via ordinal-ranking), we can also consider bargaining and auction theory. These have been developed out of game theory and are a major part of the research that considers game-theoretic problems. In retrospect, the first bargaining problem was presented by Edgeworth (1881). This described what today we would call the set of individual rational Pareto-optimal

agreements. Musgrave (1959, p. 67) says that 'a given economic agreement is Pareto-optimal if there can be no agreement which will leave someone better off without worsening the position of others'. A number of bargaining models, including the non co-operative bargaining model of Nash (1950), were developed subsequently using the non co-operative game theory as their basis. Some of these models were developed by Osborne and Rubinstein (1990), Roth (1983), Roth and Schoumaker (1983), Roth (1985), Roth, Prasnikar, et al. (1991), Camerer (1997) and Carraro, Marchiori and Sgobbi (2005).

The contribution to this discussion made by Crawford (1982, p. 607) shows that improving efficiency in bargaining outcomes does have substantial welfare gains. Brams (1994) has also posited a new concept called the 'theory of moves', which provides an alternative solution to the classical game theory framework, specifically concentrating on dynamic games. While this theory has a different way of deriving the underlying solution, it provides similar outcomes to those provided by classical game theory. To solve bargaining problems, Brams and Taylor (1996) provide a solution to the problem of splitting a fixed-size pie (i.e. decisions made by each player does not change the size of the pie). However, in many negotiating situations, the pie can change size due to the interaction (decisions) between players in a game (i.e. it depends on whether they co-operate or compete with each other). Also, if the goods or services are heterogeneous in nature this solution may not work as well, because players in a strategic interaction may prefer one type of good or service over another as they rate products comparatively. Effectively, this comes back to the question of Sen's (1977) discussion on utility and how one values goods or services against others in a bargaining situation.

The bargaining process helps value creation, destruction and distribution as each player makes decisions to improve their final payoff in the game. Hopmann (1995) takes this discussion to the next step, analysing the conceptual difference between bargaining and problem solving paradigms in game theory. This discussion is significant because it clarifies the roles of co-operation and competition in the value creation processes. According to Hopmann, bargaining is the distribution of the pie into multiple pieces through concessions provided by each player to their opponent. However, in this paradigm, each party tries to maximise their share of the pie using Brams and Taylor's (1996) concept of fair division of a fixed pie. This environment simply encourages a competitive approach to value distribution between the players. In contrast, problem-solving is a co-operative process, where each disputant will work with the other disputant in the game to resolve the outstanding issues through mutual agreement. Thus, each works to improve each player's utility and resultantly increases the size of the pie that is shared between the players in that game.

Hopmann (1995, p. 24) emphasises that problem solving results in greater flexibility, more 'frequent, efficient, equitable and durable' solutions compared to bargaining. Brams (1994) and Roth (1991) restate that game-theoretic models assume a player to be motivated by self-interest. However, the 'dual-concern model' (Pruitt and Rubin 1986, p. 29) and 'social-utility model' (Loewenstein, Thompson and Bazerman 1989, p. 426) show that players are also worried about the outcomes reached by their opponents. Reputational concerns and fairness of outcomes in a game will push the players to choose outcomes that diverge from the intuitive game theoretic paradigm (Guth and Tietz 1990, Hoffmann et al. 1998, Kahneman, Knetsch and Thaler 1986, 1990, Ochs and Roth 1989, Roth 1991, Kramer et al. 1993). Other research also supports the conclusion that players will usually divide resources equally, especially if

normative or similar contextual cues do not suggest an alternate allocation (Allison, McQueen and Schaerfl 1992, Messick 1992, Messick and Schell 1992, Bazerman and Neale 1983, Neale and Northcroft 1991, Kramer et al. 1993).

While game theory provides an explanation of strategic interaction in economic environments, behavioural economics, on the other hand, provides an application of psychology to economic thought. In recent decades, game theory and behavioural economics have been applied together in order to understand both game theoretic and psychological effects in economic interaction. Camerer (1997, 2003a, 2003b, 2004) and Benoit (2007) state that game theory outcomes in many cases by themselves may not be correct due to the human behavioural assumptions that have not been included. They state that psychological and neuroscientific theories can be used to help support game theory to provide more realistic solutions to problems, such as that of social cost (see e.g. Coase 1976, Sen 1977, Binmore 1998, Rabin 1993 with reference to Arrow and Debreu 1954, Debreu 1959, Chatterjee and Samuleson 1983, Cramton 1992, Holmstrom and Myerson 1983, Satterwaite and Williams 1989, Beaulier and Caplan 2007).

4.3 Complex systems application to economics

In the past, game theory has been applied to experiments in order to understand the concepts of selfishness, altruism and strong reciprocity, where mathematical methods have been used to model such human behaviour. Over the past two decades, however, complex systems have been developed that have become significantly more useful in performing these experiments. Social behaviour in human societies is a highly

complicated concept. As a result such behaviour can be modelled well by the mathematics of complex systems.

Complex systems are systems with nonlinear interactions and complex feedback loops, in which it is difficult to distinguish between cause and effect. It is also hard to understand a complex system as a sum of its parts due to the complex interrelationships between the different parts within the system (Rastetter et al. 1992). Holland and Miller (1991) wrote one of the first articles to identify the usefulness of complex adaptive systems for modelling economic theory. Research in the area can be found in the proceedings of the conferences on complex systems and complexity theory held at the Santa Fe Institute that are provided in Anderson et al. (1988), Arthur et al. (1997) and Blume and Durlauf (2005). Similarly, Chen (2007) provides an overview of work on computationally intelligent agents in economics and finance, which include computational intelligence and agent-based modelling and simulation.

Broadly, three research streams that relate to complex-systems in economics relate to the discussion in this thesis: ecological-economic systems, agent-based computational economics (ACE) and agent-based social simulations. Costanza et al. (1993) explain that both ecological and economic systems are complex systems and that ecological-economic systems try to emulate similarities between ecology and economic systems. The main such aspect is the application of the evolutionary paradigm. Swenson et al. (2000), for example, developed an 'artificial ecosystem selection' (p. 9110) experiment in which they analysed complex interactions from an evolutionary perspective.

Agent-based computational economics (ACE), on the other hand, is a field of research that uses computational techniques to solve economic problems. ACE research covers agent based computational learning, evolution of norms, modelling

economic networks and organisations and building computational laboratories to analyse real problems (Tesfatsion 2002, 2003).

While, agent-based social simulations are the third research stream that relates complex systems to economics. Here, Gotts et al. (2003) review the field of 'agent-based social simulations' (p. 3) and state that these models mainly relate to reciprocal altruism and the Prisoners' Dilemma problems. For example, see Bechlivanidis (2006), where an agent-based model is developed to analyse the role of prestige in a cultural evolutionary game. As prestige is considered by him as an indicator of success in society, he believes that learning from successful people decreases the cost of knowledge acquisition. On the other hand, Kim and Taber (2004) have developed an agent-based social simulation to analyse political cognition in two-person Prisoners' Dilemma games. More specific economic models have also been developed, for example, Ketelaar et al. (2007) have developed an agent-based social simulation that is labelled EMOTLAB which is used to study emotional signalling in social bargaining games.

4.4 Complex systems and strong reciprocity

Complex systems provide a computational laboratory for analysing economic behaviour. Strong reciprocity relates directly to altruistic and selfish traits in human nature. Hence, it has been important to understand strong reciprocator behaviour. Strong reciprocity, like selfishness and altruism, has been studied using computational models over the past two decades as computers have become more powerful. While mathematical models were predominantly used earlier, a number of computational models have been developed recently. Researchers have used them to analyse complex

economic behaviour including altruism, selfishness and strong reciprocity (Bowles and Gintis 2000, 2002a, 2003b, 2004, Bowles et al. 2003, Gintis 2000a, 2000b, 2000c, 2005, Gintis et al. 2003, Fehr and Fischbacher 2002, 2003, 2004a, 2004b, 2005, Fehr and Gächter 2000, 2002). Before examining such models thoroughly, however, it is prudent to say a word concerning n-person and two-person game-theoretic and complex-systems models. This will help further locate the direction of this thesis. A significant starting point is Axelrod's (1984) development of a two-person Prisoners' Dilemma game called Tit-for-Tat. In this game, an individual would co-operate with his opponent for the initial game and then mimic the strategy that his opponent followed in the previous game. Axelrod's (1984) results show that this Tit-for-Tat strategy was quite successful for this player. In contrast, Manhart (2007) developed the n-person form of Axelrod's (1984) Tit-for-Tat game and realised that the results from his model were similar to those of Axelrod (1984) except that co-operation declined quickly in n-person games compared to Axelrod's (1984) two-person game and it was hard to resolve this lack of co-operation especially in larger groups.

It is important to note that economists believe that two-person games are too simplistic and that they occur less often in the real world. As a result, the majority of the game-theoretic research undertaken in the past has been in the area of n-person games. In the research area of strong reciprocity, n-person games are used to model Prisoners' Dilemma games, while two-person games are used in modelling ultimatum and dictator games. It is also important to note that two-person games have been developed to analyse economic behaviour using complex systems. This will be evident as we discuss the key papers in this chapter that relate to the concept of strong reciprocity, which will provide a background for the 2PRIM model, that will be developed in chapter 5.

As we have discussed in chapter 3, the concept of strong reciprocity was developed by Bowles and Gintis (2000). Their model is discussed below in detail to provide a sound background relating to the research in this area. It is then followed by other models that have been developed using the Bowles and Gintis (2000) model.

4.5 Explaining the strong reciprocity model

Bowles and Gintis (2000) state that players in their model will have higher fitness when they co-operate with other players, effectively increasing their *baseline fitness* from less than *0* to $q - b > 0$, where q represents output and b represents cost/effort for doing the work (all benefits and costs are in fitness units). However, players can deceive (shirk) other players by putting in less effort if the common pool is equally shared. So, each player j could deceive σ_j fraction of the time in the game, resulting in the average level of shirking equalling:

$$\sigma = \sum_{j=1}^{n} \sigma_j / n \qquad (1)$$

Their game is modelled using a group size of n players with the overall fitness value of the group equalling $n(1- \sigma)q$. Each player's fitness value will therefore be equal to $(1- \sigma)q$. As, a result if player j deceives then the shortfall to the group will be $q\sigma_j$. The benefit of shirking to player j will equal the fitness cost of shirking that will be $b\sigma_j$ with $b(0) = b$, $b(1) = 0$, $b'(\sigma_j) < 0$ and $b''(\sigma_j) > 0$, where $q(1- \sigma) > b(\sigma_j) \ \forall \ \sigma_j \in (0,1)$. Bowles and Gintis (2000) assume that the group size n is sufficiently large, then the equation can be written as $q(1-\sigma_j)/n < b(\sigma_j)$ for $\sigma_j \in (0, 1)$. However, the group will obtain a higher output from player j's effort than the benefit player j would receive from

shirking. If player j completely shirked ($\sigma_j = 1$) in which case his fitness would increase regardless of the effort undertaken by the group.

But, if player j could be monitored and punished by other players within the group and if the fitness cost of punishing player j equalled $c > 0$, then he could be punished with a probability of $f\sigma_j$, where f represents the fraction of players in the group who are reciprocators. Bowles and Gintis (2000) explain punishment as the ostracising of the player from the game for a few periods (work alone) before letting him rejoin the game. The fitness cost to player j when he is ostracized equals $s > 0$ (an endogenous variable that is ascertained by the allocation of players that have been ostracised and those still playing the game).

Bowles and Gintis (2000) state that there are two types of individuals within their model: *reciprocators* (who are altruistic and punish shirkers that free ride with a probability equalling 1) and *self-interested* individuals (who free ride to improve their own fitness). They argue that selfish individuals free ride and there is a loss of fitness within the group as reciprocators face a cost for monitoring these self-interested individuals. In their model, the fraction f of reciprocators in the group is common knowledge, though it is not possible to differentiate reciprocators and self-interested agents at the individual level, which means that monitoring is required. Also, each player needs to be monitored with equal probability as there is no history of which individuals have been ostracized.

Therefore, the cost of working for player j, i.e. $\hat{b}(\sigma_j)$, is equal to the cost of effort and the expected cost of ostracism:

$$\hat{b}(\sigma_j) = b(\sigma_j) + sf\sigma_j \qquad (2)$$

On the other hand, self-interested individuals face a fitness loss if they get ostracized from the group that is equal to:

$$s = t_o(\phi_n(f^*) - \phi_o). \quad (3)$$

Bowles and Gintis (2000) intented to understand if self-interest is a stable equilibrium and if it can be invaded by a small fraction of reciprocators.

Initially, the game starts with a fraction ε of reciprocators in a very large population *N* of self-interested individuals. At the beginning of each period, a group of $n = \delta N$ individuals (where, δ is a very small positive number) is formed at random from the population. As, the number of individuals in the population is high, therefore the ostracised individual will spend a lot of time out of the group, as $p = n/N$ and this time can be represented by:

$$t_o = \frac{1-p}{p} = \frac{1}{\delta} - 1 \quad (4)$$

Results indicate that the population can sustain reciprocators and they increase when there is even a small fraction of reciprocators in the game. This occurs as there is a likelihood that a Nash equilibrium exists, where no agent will deceive and reciprocators have the same fitness as other individuals within the group with the group growing at a positive rate, which shows that a small fraction of reciprocators can enter a population of self-interested individuals.

While, Bowles and Gintis (2000) conclude that self-interest cannot be sustained under small values of p as λ_n is negative, which explains that individuals are shirking. As group size increases and as reciprocators increase in number, their presence in the pool also increases.

In their simulation, Bowles and Gintis (2000) set $b = c = 0.15$, $\gamma = 0.07$, $\mu = 0.08$ while assuming a baseline fitness of $\phi_o = 0.02$. Also, adjusting the productivity of effort q till population size is an approximate constant when equilibrium is reached. This provides a value of $q = 0.19$. They find that a fraction of reciprocators in the groups $f^* = 70\%$, the shirking level of self-interested agents is $\sigma_n(f^*) = 10\%$, the fitness cost of ostracism is $s^* = 0.29$ and the average time in the solitary pool is $t_o = 5.38$ periods. Thus, the fitness of the reciprocators is:

$$p_r \pi_r + (1 - p_r)\phi_o = 0.00 \qquad (5)$$

Further, the fitness of the non-reciprocators can be seen as:

$$p_n \phi_n + (1 - p_r)\phi_o = 0.01 \qquad (6)$$

This result should also be equal to zero, however Bowles and Gintis (2000) state that a rounding error has occurred, primarily as the equilibrium values of f, s, λ_n and λ_g are only accurate to two decimal places. This leads us to the next section to review subsequent research that has been developed in the area of strong reciprocity based on the Bowles and Gintis (2000) model.

4.6 Work in the tradition of Bowles and Gintis

The concept of strong reciprocity as developed by Bowles and Gintis (2000) and Gintis (2000a, 2000b, 2000c) is differentiated from that of reciprocal altruism. Bowles and Gintis (2000) also provided the starting point for strong reciprocity research using complex systems.

Boyd et al. (2003) subsequently built on the strong-reciprocity approach provided by Bowles and Gintis (2000). They state that it is possible for co-operation to exist in small groups of unrelated people, due to the fact that people even punish non co-operators in one-shot experiments which support co-operation. However, they question if co-operation can exist in larger groups as individuals do not get the direct benefit of the co-operative behaviour. So, they develop an evolutionary simulation that has 127 groups of defectors and 1 group of altruistic punishers with this game played for 2000 time periods. Some variables in this model are: the cost of co-operation and cost of competition both having a value of 0.2, cost of being punished being 0.8, a migration rate equal to 0.001 and a mutation rate equal to 0.01. Their model also had an average extinction rate which is in line with the cultural extinction rate in small societies equalling 0.0075.

Their results confirmed that altruistic punishment and altruistic co-operation can exist due to group selection and co-operation cannot exist without punishment in large groups. They also find that co-operation falls significantly with an increase in the migration rate and the cost of punishment. They state that without punishment co-operation can exist in small groups only. As, the cost of monitoring defectors reduces, there is an increase in co-operation.

While Boyd et al. (2003) analyse strong reciprocity and the differences between altruistic co-operation and altruistic punishment. Fehr and Fischbacher (2003) develop a model to analyse strong reciprocity and they find that cultural evolution could contribute significantly to develop this understanding. Bowles and Gintis (2004) develop on their previous paper (Bowles and Gintis 2000) to incorporate heterogeneous populations using an n-person Prisoners' Dilemma game, in their simulation, they have three types of agents, strong reciprocators (second-order altruistic punishers), selfish

and purely altruistic individuals. Results provided by Bowles and Gintis (2004) explain that strong-reciprocity most likely existed even 100,000 years ago, that all three of these behavioural types possibly survived in that environment and these results did not require any individuals to be genetically related.

Similarly, Gächter and Fehr (2004), Mendes (2004), Panchanathan and Boyd (2004), Sethi and Somanathan (2004' 2005), Fehr and Fischbacher (2004a, 2004b, 2005, 2006), Eriksson and Lindgren (2005), Calderon and Zamara (2006), Nakamaru and Iwasa (2006), Fehr and Gintis (2007), Fehr and Schneider (2007), Bowles and Hwang (2008), Carpenter et al. (2008) Gintis et al. (2008) and Tucker and Ferson (2008) have developed strong-reciprocity research using complex systems. These papers have considered strong reciprocity to be second-order altruism, where altruistic individuals punish selfish individuals in order to increase the co-operation within the game. Eldakar et al. (2007) and Eldakar and Wilson (2008) present another model, one in which strong reciprocators are not second-order altruistic. Instead they model first order selfish individuals as being second-order altruistic because they punish other selfish individuals to improve their likely payoff in the future.

Eldakar et al. (2007) and Eldakar and Wilson (2008) develop an n-person evolutionary Prisoners' Dilemma game to expand the possible models in which strong-reciprocity is effective. Each individual is assigned an altruistic and punishment trait with values between 0-1 in increments of 0.1. Members of each group play multiple rounds of a two-phase game. In the first phase, each individual is provided with a given endowment and is allowed to contribute to a common fund that is doubled at the end of phase one with the common fund proceeds distributed equally to each individual regardless of their contribution to the common fund. Payoff for each individual can be calculated as follows:

$$pay_i = E(1-A_i) + \frac{2E(\sum_{j=1}^{N} A_j)}{N} \quad (7)$$

Here, $E(1-A_i)$ is the individual payoff that has been withheld by the selfish individual (i.e. this is the amount that was not contributed to the common fund by the selfish individual).

While, $\frac{2E(\sum_{j=1}^{N} A_j)}{N}$ is the payoff received from the common fund. This payment is an equal distribution provided to all individuals in the game, who did or didn't contribute to the common fund.

In Phase two, funds can be contributed by individuals to find and punish the biggest cheater. The punished cheater will be permanently ostracised from the game and he will be replaced with a new individual at the beginning of the next game drawn randomly from the same population as the original members. Eldakar et al. (2007) state that even if replacements do not play the same number of rounds as the original players depending on how they play the remaining rounds, they still contribute to the fitness differential in the total population. The amount that is invested in punishing this individual is based on three factors:

$$punC_i = P_i \frac{(\sum_{j=1, J\neq i}^{N-1} 1-A_j)}{N-1} C \quad (8)$$

Where, P_i = individual's static punishment trait

$\frac{(\sum_{j=1,J\neq i}^{N-1} 1 - A_j)}{N-1}$ = average amount of cheating that took place among other members of the group

C = amount required to detect the biggest cheater with certainty

As a maximum of two individuals can be removed from the group at the end of each round, being the biggest cheater or the second biggest cheater. The probability that the biggest cheater cannot be caught:

$$esc_i = 1 - P_i \frac{(\sum_{j=1,J\neq i}^{N-1} 1 - A_j)}{N-1} \quad (9)$$

The probability that the biggest cheater is removed from the game can be seen as:

$$rem_{all} = (1 - \prod_{i-1}^{n-1} esc_i)D \quad (10)$$

Here, the biggest cheater is not included in the above equation. D represents the likelihood that the biggest cheater will be removed once he has been found. In the case that $D = 1$ it means that the cheater can be found and isolated from the game and when $D = 0$ it means that the cheater cannot be isolated from the game even if he is found. Nonetheless, while the group is looking for the biggest cheater, this cheater can work towards identifying the second biggest cheater, so that he does not get isolated from the game.

Eldakar et al.'s (2007) results show that equilibrium is attained between altruistic non-punishers and selfish punishers when there are an equal number of altruistic and selfish individuals with coupled oscillations occurring between these traits

in the shorter time horizon. However, the same result is found, though in a longer time span, when the population consists of only selfish non-punishers with altruistic/selfishness traits equal to zero. Eldakar et al. (2007, p. 201) explain how selfish punishment works stating that:

> To see how selfish punishment promotes the evolution of altruism, consider a single selfish punisher in a given group. By expelling the most selfish individuals, which are replaced by randomly chosen members of the total population, the punisher increases the average degree of altruism within the group. Altruists now benefit from each other and the selfish punisher recovers the cost of punishment by exploiting the altruists during subsequent rounds.

In order for this to occur there need to be enough altruists in the group. This can only happen in the latter case when the mutation rate is greater than 10^{-4} that results in a selection-mutation balance of approximately seven per cent of the population with the altruism trait greater than zero that provides for sufficient concentration of altruists in order for punishment to occur. We also need to note that if the mutation rate is below 10^{-4} or the cost of punishment is sufficiently high, then altruism will not evolve from the start and it will take substantially longer for such an equilibrium to be achieved.

Eldakar et al. (2007) also find that altruism and punishment start waning as the cost of punishment increases, though it does still remain at low levels when the cost of punishment is high. Thus, they believe that the concept of selfish punishment is even more relevant when punishing others is costly, specifically as selfish punishers can recoup these costs by exploiting altruists and that altruistic punishers do not have the same amount of resources as the selfish punishers to punish selfish individuals. The relationship between altruism and punishment is low when the group size is less than

seven individuals, but it increases as the group size goes above this limit. As the group size increases, round length becomes a more important factor in the game, as only two individuals can be eliminated from each game making the other individuals less likely to be eliminated. Also, they find that punishment has minimal impact in maintaining altruism in a single round, as a result requiring the round length to be increased for any impact to be seen.

I have reviewed the literature on Bowles and Gintis (2000), Boyd et al. (2003), Eldakar et al. (2007) and Eldakar and Wilson (2008) in this chapter. It is important to remind the reader again that this thesis intents to analyse the concept of one-shot or non-repeated (random) interactions. The literature review in chapters 1 – 3 has shown that one-shot interactions is an important research area. However, prior to developing the two-person random interaction model (2PRIM), we need to answer how this model extends the learning from the papers reviewed in this chapter, i.e. how is the work in this thesis an original contribution.

In order to analyse if this thesis has made an original contribution, while building on existing literature, we first need to specify a particular kind of everyday random interaction, those involving two people only (or 'dyadic'). This thesis develops a computer simulation model, the two-person random interaction model (2PRIM), to assist in understanding such interactions to model the respective effects of selfishness, altruism and strong reciprocity on the consequences of two-person random interactions.

The first contribution of this thesis therefore is that it models a significant problem that scholars seek to understand, namely random, one-shot or non-repeated human interactions. This aligns with the literature review in the thesis up to this point. This problem is one we face in our everyday lives in our communities and, if we travel, beyond them. We regularly engage in random interactions in our working, professional

and leisure activities. As Silk has pointed out (2005, pp. 63-4), such interactions are of evolutionary significance.

Secondly, the thesis contributes originally to knowledge by modelling a specific set of random interactions, namely those involving two persons, continuous-trait attributes and strong reciprocity. This model allows for those behaviours or traits to be represented randomly along a continuum. While other researchers address similar random or one-shot interactions, none seem to combine these three aspects in modelled interactions, i.e. as two-person, as continuous-trait and as involving strong reciprocity.

The closest parallels are those of Eldakar et al. (2007) and Eldakar and Wilson (2008). The former models strong reciprocity in N-person interactions, with traits allocated 'that initially vary uniformly ... between 0 and 1 at 0.1 increments' (2007, p. 199). However, N is always greater than two. The latter study allocates traits or behaviours for selfishness, altruism and strong reciprocity as 'pure strategies', which is to say that members of a group of N are either pure altruists or purely selfish and either strong reciprocators or not (2008, p. 6982). Moreover earlier work on co-operation using two-person models relies on 'non-random interactions or guarded cooperation' (2008, p. 6982, citing Axelrod 1984, Hamilton 1964, 1975, Axelrod and Hamilton 1981, Maynard Smith 1982) and does not consider strong reciprocity. Bowles and Gintis (2000) and Boyd et al. (2003) on the other hand have considered the values of selfishness and altruism in either the values of 0 or 1, i.e. an individual could be either totally selfish or altruistic.

Thirdly, the thesis creates an original game-theoretic computational model (2PRIM) in order to contribute to our understanding of the individual, social and evolutionary consequences of everyday two-person random interactions in defined

circumstances. While, a substantial amount of literature has been reviewed, there is no instance where one-shot random interactions between two strangers has been analysed.

Fourthly, the thesis contributes originally to knowledge by developing the foregoing contributions within the context of a theoretical discussion of the literature – including the work of Adam Smith – of this emerging multidisciplinary field of research. Indeed it is only within the context of that discussion that the questions posed by this thesis can be understood. It is only within that context, too, that the models presented in this thesis make sense. There has been specific discussion on literature related to Adam Smith in chapter 2, nonetheless subsequent chapters have analysed the concepts of selfishness, altruism and strong reciprocity from a diverse set of research areas – viewing these concepts from the context of behavioural economics, neuroeconomics, socio-biology, game theory and complex systems.

The final contribution of this thesis is that the 2PRIM shows that pay-offs for players can be modified through changes in the return on selfishness, common good or the cost of competition, which can be a substitute to applying strong reciprocity within random two-person interactions. Previous literature, for example, Gintis (2000a, 2000b, 2000c and 2007), Fehr and Gächter (2007) have shown that strong reciprocity is important. However, they have not explained that return on selfishness, common good and cost of competition can be substituted for strong reciprocity in order to manage the level of selfishness and altruism within a two-person random interaction. The contribution chapters 5 and 6 as a result build on the existing literature that has been reviewed in this thesis in the earlier chapters.

4.7 Conclusion

This chapter has reviewed research in the area of game theory, before discussing the inter-linkages between game theory, behavioural economics and social cost. It then provided a background to three related complex systems streams that help resolve economic and social problems, which are: ecological-economic systems, agent-based computational economics and agent-based social simulations. After this review, this chapter analysed how complex systems methods could be used to model the concept of strong reciprocity. More specifically this chapter examined the complex-systems models in game theory and strong reciprocity that will lead this discussion into the development of the model in the next chapter. The strong reciprocity and complex systems models that were analysed were: Axelrod's (1984) model was discussed briefly, but the three models, by Bowles and Gintis (2000), Boyd et al. (2003) and Eldakar et al. (2007), were discussed in depth. Other strong reciprocity models were also reviewed to support this discussion. This analysis provides a sound basis for the discussion of the 2PRIM model developed in the next chapter.

CHAPTER 5

Analysing selfishness and altruism in two-person random interactions

5.1 Introduction

Previous chapters have discussed the concepts of altruism and selfishness, indicating that humans portray both these emotions. Some people are more altruistic than selfish and vice versa. While selfishness helps maximise individual fitness or utility, it reduces the overall fitness or utility of the group. This can result in punishment by others. Individuals can also punish others whom they perceive to be selfish or unfair in one-off interactions, as shown in experiments involving the ultimatum game. In everyday life, however, humans come across many repeated interactions, one-off (random) two-person interactions and N-person interactions. Out of all these transactions, a significant portion of our time is spent in two-person random interactions with strangers. That is, we will engage with a person whom we have not met and likely will not meet again in our lifetime and about whom we know very little prior to this interaction. So, how might we interact in such transactions? Would it pay to be selfish or altruistic in such two-person random interactions with strangers?

This last question, of course, is impossible to answer in the abstract. We would need to know our own predisposition towards selfishness and altruism, whether and how much we would be disposed to punish and the prevailing rules of the game (or interaction). Would we adopt a fitness-enhancing (or maximising) or a different approach? This would also have to be known, and we would have to know it about the stranger, too, if we hope to answer whether it would pay to be selfish or altruistic in two-person random interactions with strangers. So, while it does not make a lot of

sense to ask the above question in the abstract, asking it does focus our attention to the necessary underlying conditions that might render the question capable of eliciting a coherent answer. This, in turn, illuminates the necessary caveats that apply to an answer when any of the above conditions are absent.

For example, in the absence of complete information, we might substitute some probabilistic information. In this case, the answer, too, will necessarily be probabilistic. In the absence of any knowledge at all, including probabilistic knowledge, any answer to the question whether it is 'better' to be altruistic or selfish is necessarily a guess, because the result is indeterminate.

All of the above must be stated at the outset, clearly and unambiguously. Why? Because the following model of selfishness, altruism and strong reciprocity creates the rules of the game and sets its limits. In particular, as touched upon in the preceding chapters, the 2PRIM model developed in this and the next chapter sets the following general determinate evolutionary game-theoretic conditions:

1. A prison's dilemma framework, in the form of two equations, that sets the structure of interaction – i.e. the dimensions and limits of possible outcomes;
2. A probabilistic allocation of altruistic, selfish and strong reciprocity propensities to the two randomly-selected interacting individuals; and
3. A method of random selection and elimination of interacting individuals that simulates evolution.

In consequence of these general determinate conditions the 2PRIM model has been created in accordance with the prevailing game-theoretic approach, in which selfishness is usually seen as a dominant strategy. It should be interesting, however, to see under

what conditions selfishness or altruism will dominate in a two-person random interactional model (2PRIM), where random interactions take place between unrelated individuals (i.e. strangers). That is, under which conditions selfish individuals will gain higher utility or fitness in games at the expense of the altruistic. If altruism is a common (or public) good and, as a result, utility is highest when two altruistic individuals meet, under what circumstances does this behavioural trait dominate over selfishness? Consider a necessary meeting with a stranger on the street whom you have never seen before. If you must interact, how would you engage in an undertaking with this person? Would you act selfishly or altruistically, and how would they react to your actions? In this chapter the former issue is considered, while the latter issues will be considered in chapter 6.

Chapters 2 -4 have analysed selfishness, altruism, strong reciprocity, evolutionary game theory and complex systems, which provide the basis for developing the 2PRIM model. Chapter 4 explained how Axelrod (1984) had developed a two-person model that included interactions between individuals using a repeated tit-for-tat strategy. In Axelrod's game, a player will follow the strategy that was played by their opponent in the previous round. However, in 2PRIM, individuals are selected using a simple random process. They compete with each other in a one-off two-person Prisoners' Dilemma game based on their levels of selfishness and altruism that are assigned at random. Models developed in Bowles and Gintis (2000), Eldakar et al. (2007), Guth and Yaari (1992) and Boyd et al. (2003) have been used as a guide to develop 2PRIM. Adam Smith's concepts of altruism and selfishness are primarily used in developing the human behavioural traits that underlie the construct of 2PRIM in this chapter.

Conceptually, this thesis intends to present the following ideas that will help extend the existing literature to better understand human behaviour in two-person everyday random interactions. The model in this thesis intends:

1. *To model how, in an evolutionary sense, an individual's selfishness traits result in a higher level of individual utility in everyday random interactions.* 2PRIM is used to represent a Prisoners' Dilemma form of one-on-one random interactions that people have with each other in everyday life.
2. *To understand the change in equilibrium within the two-person game in the presence of Adam Smith's impartial spectator (i.e. the feeling of guilt). This 'guilt' parameter adjusts the amount of benefit derived by a selfish individual depending on this person's level of guilt.*
3. *To observe how changes in the selfish and altruistic investment, defined below, have an impact on the levels of selfishness and utility in this model. This thesis analyses how these two factors affect the individual's payoff. This concept is developed further in Chapter 6 to understand if changing these two factors can have a strong impact on group or total utility compared with changes in punishment (strong reciprocity).*
4. *To observe how competition between selfish individuals, which comes at a cost in the model, will change the evolutionary properties of the model. Previous research has not examined specifically how the return on the common pool and selfish investments, (defined below) can affect the level of altruism in groups. However, there is some analysis undertaken of the cost of competition in n-person strong-reciprocity games. Analysis in chapter 6 will also compare if changes in the cost of*

competition factor has a greater affect compared to changes in the level of punishment.

Selfishness and altruism are behaviour traits that individuals possess in 2PRIM. However, it is important that there is evolution in this model for equilibrium to be established as individuals with the less suitable trait obtain lower utility and are removed at the end of each game. As individuals are removed they are replaced with new individuals with a random level of selfish/altruistic (A-S) trait. The dynamics of the model assists in evolving the population and establishing equilibrium.

This chapter will provide a brief review of some of the relevant literature relating to this model. The third section will then outline the 2PRIM model systematically. Subsequent sections will analyse the preliminary results obtained from this model, and the final section will summarise the key learning from this chapter.

5.2 From Adam Smith to game theory and the 2PRIM model

Adam Smith has discussed the concepts of selfishness and beneficence in *An Inquiry into the Nature and Causes of the Wealth of Nations* (Smith 1776) and *The Theory of Moral Sentiments* (Smith 1790) respectively. Ashraf et al. (2005), Montes (2003) and Evensky (2005) have stated that Adam Smith's concepts of selfishness and beneficence are complementary in nature and that every individual possesses the behavioural traits of selfishness and altruism. This leads us towards the concept of strong reciprocity, which resonates with Smith's theory. Strong reciprocity as a concept was introduced specifically in game theory by Bowles and Gintis (2000). It describes second-order altruistic behaviour, where an altruist will punish selfish individuals in a game in order to increase co-operation within a group. Bowles and Gintis (2000) and others explain

experimental evidence that strong reciprocity exists among unrelated individuals, and this behaviour cannot be explained by theories of kin selection, reciprocal altruism, costly-signalling and indirect reciprocity.

Interestingly something akin to the following framework is foreshadowed in one aspect of Adam Smith's discussion of utility. Utility, according to Smith, was of subordinate status morally to virtue. Virtue saw individuals exercise proper beneficence (altruism) and justice in their dealings with others and proper prudence (self-interest) in pursuing their own interests. Recall from chapter 2 that the virtue of self-command was also necessary to keep in check both the unruly passions and, in particular, any tendency to improper, excessive self-interest. While it was proper to 'strain every nerve and every muscle' to get ahead, it was wrong to 'jostle' or 'throw down' a competitor in order to gain an unfair advantage (Smith 1790, p. 83, pp. 137-8).

Utility nonetheless had a role socially. As in the Prisoners' Dilemma, society was better off when its members acted altruistically or co-operatively:

> It is thus that man, who can subsist only in society, was fitted by nature to that situation for which he was made. All the members of human society stand in need of each other's assistance, and are likewise exposed to mutual injuries. Where the necessary assistance is reciprocally afforded from love, from gratitude, from friendship, and esteem, the society flourishes and is happy. All the different members of it are bound together by the agreeable bands of love and affection, and are, as it were, drawn to one common centre of mutual good offices. (Smith 1790, p. 85)

Yet, even if society's co-operation were merely a matter of self-interested or prudent or utilitarian exchange, it might still stay together (as when prisoners' agreement not to rat holds):

> But though the necessary assistance should not be afforded from such generous and disinterested motives, though among the different members of the society there should be no mutual love and affection, the society, though less happy and agreeable, will not necessarily be dissolved. Society may subsist among different men, as among different merchants, from a sense of its utility, without any mutual love or affection; and though no man in it should owe any obligation, or be bound in gratitude to any other, it may still be upheld by a mercenary exchange of good offices according to an agreed valuation. (1790, p. 85-6)

However, once society starts to act like prisoners who rat, a race to the bottom begins. It is like the eventual equilibrium of the Prisoners' Dilemma game: the lose-lose outcome caused by one prisoner's selfishness being replicated by the other. It is the outcome Smith envisaged in his references to selfishness (excessive self-interest) in wanting to 'jostle' or 'throw down' competitors unfairly. As Smith puts it:

> Society, however, cannot subsist among those who are at all times ready to hurt and injure one another. The moment that injury begins, the moment that mutual resentment and animosity take place, all the bands of it are broke asunder, and the different members of which it consisted are, as it were, dissipated and scattered abroad by the violence and opposition of their discordant affections. If there is any society among robbers and murderers, they must at least, according to the trite observation, abstain from robbing and murdering one another. Beneficence, therefore, is less essential to the existence of society than justice. Society may subsist, though not in the most comfortable state, without beneficence; but the prevalence of injustice must utterly destroy it. (1790, p. 86)

Recall also from chapter 4 that Eldakar and Wilson (2008) contrive two similar pure strategies. Their focus is the roles of selfishness and altruism in evolution, with an emphasis on biological fitness. They adapt the public goods game to this end. As they explain:

> ... emulates an experimental economics game in which each member of a group is provided an endowment, b, that can be kept or invested in a public good. The combined investment in the public good is multiplied by a factor, m, and distributed equally to everyone in the group. The total payoff of each individual (the proportion of the endowment kept for oneself plus one's share of the public good) is assumed to be linearly related to fitness. This scenario can easily be related to biological situations, such as investing effort in a hunt in which everything captured will be shared. The model considers the two pure strategies of investing all (altruist) or none (selfish) of one's endowment ... It is clear that in the absence of punishment, selfish individuals always have the highest fitness within the group because all group members obtain an equal share of the contributions from the altruists, yet selfish individuals keep rather than donate b. Therefore, an altruist would obtain a greater fitness by switching to the selfish non-contributing alternative, resulting in a selfish gain of b minus the now share of the reduced group payoff caused by the loss of a single altruist ... However, the group has the highest fitness when everyone is an altruist, resulting in the classic Prisoners' Dilemma situation. (2008, p. 6982)

It is from models such as this that selfishness and altruism are discussed in terms of investment and effort. Altruism, as such, is an investment, the deployment of effort. In such senses it is a cost by definition. Selfishness, on the other hand, is seen as not investing and not exerting effort. It is an opportunity gain, by definition. (Note, that labour as effort is usually assumed to be a disutility, a decrement to fitness and a cost. It is a questionable assumption but one that I will not pursue in this thesis.)

The 2PRIM model endeavours to consolidate the above conceptions in a basic equation. As noted earlier this equation sets out the structure of interaction between the two random actors who engage in a one-shot interaction. For players 1 and 2 the basic forms of the equation are:

$$\frac{A_1 + A_2}{2} \cdot CGfactor + S_1 \cdot SFfactor - (S_1 + S_2) \cdot Cfactor \quad (1)$$

$$\frac{A_1 + A_2}{2} \cdot CGfactor + S_2 \cdot SFfactor - (S_1 + S_2) \cdot Cfactor \quad (2)$$

In the 2PRIM basic equation:

1. A_1, A_2 represent the altruistic traits of players 1 and 2
2. S_1, S_2 represent the selfishness traits of players 1 and 2
3. $1 - A_1 = S_1$ on a continuum from 0 to 1
4. $1 - A_2 = S_2$ on a continuum from 0 to 1
5. *CGfactor* is a factor representing returns to the common good from the cost of altruism
6. *SFfactor* is a factor representing returns to selfishness or its opportunity gain
7. *Cfactor* is a factor representing the costs of selfishness

Note that 'trait' is defined in the model in standard fashion:

A characteristic or quality distinguishing a person or (less commonly) a thing, especially a more or less consistent pattern of behaviour that a person possessing the characteristic would be likely to display in relevant circumstances. (Colman 2009)

The first term of equations (1) and (2) is:

$$[(A_1 + A_2)/2].CGfactor \quad (3)$$

The common good factor (*CGfactor*) equates to the return on the common good (altruistic) investment ($A_1 + A_2$), which is the cost of both players being altruistic to some degree. As the first term is the return both individuals get for investing in the common pool, it is shared equally (halved) between the players. It is a linear function of the total altruism of both. It might be thought of as being akin to the public good derived from donating to charity or paying taxes (ignoring complications associated with different capacities to pay). A more selfish individual, who does not donate to charity and/or who evades taxes, also benefits equally from the common pool.

The second term of equations (1) and (2) is:

$$S_1.SFfactor \text{ or } S_2.SFfactor \quad (4)$$

The selfish factor (*SFfactor*) equates to the return on the selfish investment (i.e. the return on S_1 or S_2). Note that 'selfish investment' means precisely *not* investing in the common good. For example, if I do not contribute to charity or do not pay my taxes (governmental taxes being a social redistribution of income between the citizens of a country), then by definition I save. In this way, I am not sharing some of my income and I act selfishly. In Eldakar's and Wilson's language, $S_1.SFfactor$ represents the opportunity gain of player 1 by not exerting, investing or incurring a fitness cost due to effort.

The third term of equations (1) and (2) seeks to capture the idea described by Adam Smith above, which is selfishness (jostling or throwing down) itself comes at a

cost in the form of negative public good (or public bad). It represents the social cost of competition manifested in social breakdown, for example. The third term is:

$$-(S_1 + S_2).Cfactor \quad (5)$$

The third factor, the cost of competition factor (*Cfactor*), explains the cost of destructive selfishness (i.e. the negative return to total selfishness, $S_1 + S_2$). This occurs when two selfish individuals confront each other. In this case, both the selfish individuals want to do better and compete with each other, in effect, in avoidance behaviours. They end up eroding the benefits each of them would otherwise have achieved if they had met an altruistic individual. In the 2PRIM model, selfish individuals can take advantage of altruistic individuals, as selfish individuals gain utility from the common pool investment even if they do not contribute to the common pool. They obtain higher utility as they obtain half of the common pool. For example, if two individuals bid to purchase a block of land, if both keep bidding higher to purchase this block then both will have to pay more to acquire this block as the price will increase. If one of them ends up entering the winning bid, then he or she could be worse off than their opponent. Regardless, they have created disutility for each other, rather than if they had met an altruist who might not have bid up the price against them. The winning bidder may have purchased that block of land at a much cheaper price.

Above, I have outlined the sense in which 2PRIM uses the terms in order to recreate the Prisoners' Dilemma framework. I have used the language of Eldakar and Wilson (2008) to illustrate a case, to which it might apply. However, before proceeding to the evolutionary results of the 2PRIM model in detail, it would be useful to set out its properties. That is, in addition to the structure (i.e. the equation), it is important that we have some understanding of the dimensions and the limits of its possible outcomes when two individuals meet.

First, consider how the model makes selfishness the individually dominant strategy in all two-person interactions, which is the standard approach in Prisoners' Dilemma frameworks. If we subtract equation (2) from equation (1), which is to say that we calculate the difference in the utility (fitness, payoff) between the two randomly-selected players, we get:

$$U_1 - U_2 = (S_1 - S_2).SFfactor \quad (6)$$

Therefore, if $S_1 > S_2$, then $U_1 > U_2$, which is to say that the more selfish player will always have a greater return in the model. Note that this, of course, will act to determine subsequent results. Our attention will be on how it does so in an evolutionary sense. Note also that equation (6) means precisely that the difference in the utility (fitness, payoff) between players 1 and 2 is determined exclusively in the advantage that the more selfish exacts by 'jostling' or 'throwing down' the less selfish – i.e. by the greater the amount of his or her selfishness.

Second, consider total utility (fitness, payoff). It will be the sum of the outcomes of players 1 and 2 (i.e. the sum of equations (1) and (2)):

$$(A_1 + A_2).CGfactor + (S_1 + S_2).SFfactor - 2.(S_1 + S_2).Cfactor \quad (7)$$

Now, given that $S_{1,2}$ and $A_{1,2}$ exist on a continuum from 0 to 1, when $A_{1,2} = 1$, $S_{1,2} = 0$. Hence $\sum U_{1,2} = 2.CGfactor$. In contrast, when $S_{1,2} = 1$, $A_{1,2} = 0$, and:

$$\sum U_{1,2} = (S_1 + S_2).SFfactor - 2.(S_1 + S_2).Cfactor = 2.(SFfactor - 2.Cfactor) \quad (8)$$

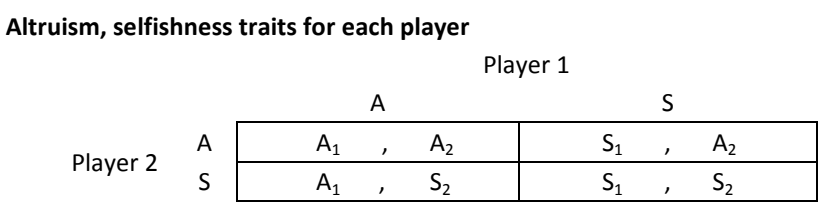
Thus, to meet another Prisoners' Dilemma constraint, which is that total utility (fitness, payoff) is greater when both players are altruistic, it must be so that:

$$2.CGfactor > 2.(SFfactor - 2.Cfactor) \rightarrow CGfactor > (SFfactor - 2.Cfactor) \qquad (9)$$

This is to say that the returns to altruism (common good) must be greater than the difference between the return to selfishness and two times the costs of selfishness. If *Cfactor* = 0, for instance, it must be so that *CGfactor* > *SFfactor*. In other words, an implicit constraint on values chosen for *CGfactor*, *SFfactor* and *Cfactor* in the next section is that they do not violate this prisoner's-dilemma condition.

Third, consider the pre-determined results of each two-person interaction that reflect total altruism, total selfishness and mixed 'strategies'. With given values for *CGfactor*, *SFfactor* and *Cfactor* the nominal maximum and minimum values for altruism and selfishness shown in figure 5.1 below will give the maximum and minimum results for each type of interaction. Figure 5.1 also shows these results (i.e. utilities, fitness or payoffs) for each player and in total. Purely for convenience *CGfactor*, *SFfactor* and *Cfactor* are set at 1 (constant returns to scale). It is easy from figure 5.1 to infer the following important characteristics of the model: maximum and minimum possible dimensions for utility, fitness or payoff are determined entirely by the values chosen for *CGfactor*, *SFfactor* and *Cfactor*. This point is reinforced in figure 5.2, which uses the same framework to illustrate formulae for utilities when altruism and selfishness are at their maxima and minima.

Figure 5.1 2PRIM's structure and dimensions for two-person interactions (maxima and minima)



Nominal altruism, selfishness traits for each player

		Player 1 A	Player 1 S
Player 2	A	1.0 , 1.0	1.0 , 1.0
	S	1.0 , 1.0	1.0 , 1.0

Utility, fitness, payoff for each player

		Player 1 A	Player 1 S
Player 2	A	1.00 , 1.00	0.50 , -0.50
	S	-0.50 , 0.50	-1.00 , -1.00

CGF = 1.00
SF = 1.00
CF = 1.00

Total utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	2.00	0.00
	S	0.00	-2.00

CGF = 1.00
SF = 1.00
CF = 1.00

Figure 5.2 2PRIM's formulae for utility from two-person interactions (maxima and minima)

Altruism, selfishness traits for each player

		Player 1 A	Player 1 S
Player 2	A	A_1 , A_2	S_1 , A_2
	S	A_1 , S_2	S_1 , S_2

Nominal altruism, selfishness traits for each player

		Player 1 A	Player 1 S
Player 2	A	1.0 , 1.0	1.0 , 1.0
	S	1.0 , 1.0	1.0 , 1.0

Utility, fitness, payoff for each player

		Player 1 A	Player 1 S
Player 2	A	CGF , CGF	CGF/2 + SF – CF , CGF/2 – CF
	S	CGF/2 – CF , CGF/2 + SF – CF	SF – 2CF , SF – 2CF

This third set of conditions helps us to derive an important constraint of the Prisoners' Dilemma framework (one which will be violated later). That is, there must

be an incentive for a given player to rat. In terms of figure 5.2, it means that *CGfactor* (the altruistic option) must always be less than *CGfactor* /2 + *SFfactor* – *Cfactor* (the selfish alternative). In other words, this condition may be simplified as:

$$2.(SFfactor - Cfactor) > CGfactor \quad (10)$$

Or, taking into account points one and two above, the Prisoners' Dilemma conditions are met by the 2PRIM model when (for player 1):

$$S_1 > S_2 \text{ and } 2.(SFfactor - Cfactor) > CGfactor > (SFfactor - 2.Cfactor) \quad (11)$$

Having now given an account of the first of 2PRIM's determinate conditions (structure and dimensions of each two-person interaction), this chapter turns in the next section to the remaining two determinate conditions of the model. These are: its process of random selection and its evolutionary characteristics. Section 5.3 also adds more detail to the discussion, as well as presenting preliminary outcomes when punishment (strong reciprocity) is not involved.

5.3 The two-person random interaction model (2PRIM) in detail

In 2PRIM, an individual can have an altruism/selfishness (A-S) trait value anywhere from 0.0000 to 1.0000 (assigned to 4 decimal places for accuracy). This is defined as the A-S continuum. Different values along the A-S continuum help provide for variety in individual behaviour, which is consistent in representing the individuality, variety and uniqueness in human behaviour in the real world. In the past most researchers, for example Bowles and Gintis (2000) and Boyd et al. (2003), have only considered specific discrete A-S trait values of either 0 or 1. However, Eldakar et al. (2007) use

discrete values of altruism and punishment between 0.00 and 1.00 with uniform increments of 0.10. They do not consider any value between these limits and do not use the A-S continuum concept. Eldakar et al. (2007) and Eldakar and Wilson (2008) consider four types of individuals in their game, which were initially identified in Nakamura and Iwasa (2006). These individual types are: selfish punishers, selfish non-punishers, altruistic punishers and altruistic non-punishers.

These four individual types comprise two trait continua: altruism (A) – selfishness (S) and non-punisher (NP) – punisher (P). This chapter only discusses the altruism (A) – selfishness (S) continuum. However, in Chapter 6, the second continuum of non-punisher (NP) – punisher (P) will also be discussed. The four individual types identified by Nakamura and Iwasa (2006) are the combination of the four extremes of these two continua. These four individual types have the following altruistic-selfish (A-S) trait values: selfish punisher and selfish non-punisher have an A-S value equal to 1.0000 and altruistic punisher and altruistic non punisher have an A-S trait value of 0.0000.

The altruistic-selfish behaviour of each individual in 2PRIM can be considered to be a point on this continuum in the range of 0.0000 – 1.0000, which, as noted above, means that each person's behaviour is a combination of altruism and selfishness. Based on the A-S trait value, each individual invests in either the common pool (altruistic investment) and/or the selfish investment. For example, if an individual has an A-S trait value of 0.2500, then they will invest 75.00 per cent in the selfish investment and 25.00 per cent in the altruistic investment. In addition to the altruistic-selfish continuum, there are three other important variables in 2PRIM. These three factors are: the common good factor (*CGfactor*), selfish factor (*SFfactor*) and cost of competition (*Cfactor*), which were discussed above. The evolutionary dynamics of 2PRIM have been developed and simulated in *MATLAB R2008b*. 2PRIM follows this process:

1. a pool of 1000 individuals are selected from a pool of N individuals at random;
2. the A-S traits are assigned to these individuals using a simple random process;
3. each experiment comprises 1000 games;
4. each game comprises 100,000 rounds;
5. two individuals are chosen at random for each round;
6. at the end of the round the utility (payoff, fitness) of each of these players will be calculated based on their altruism-selfishness (A-S) trait value using the equation below (for players 1 and 2):

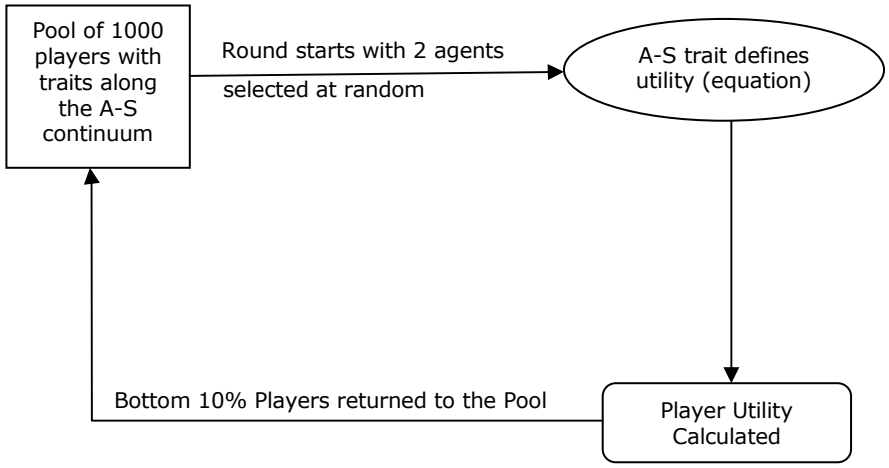
$$\frac{A_1 + A_2}{2} \cdot CGfactor + S_1 \cdot SFfactor - (S_1 + S_2) \cdot Cfactor$$

$$\frac{A_1 + A_2}{2} \cdot CGfactor + S_2 \cdot SFfactor - (S_1 + S_2) \cdot Cfactor$$

7. both are sent back into the pool of 1000 individuals;
8. at the end of each game the 1000 individuals are sorted from highest to lowest utility;
9. the bottom 10 per cent of individuals are eliminated and replaced by new individuals that are selected at random from the pool of N individuals;
10. these new players will have their A-S trait values assigned at random;
11. the utility values of each player (in the pool of 1000 individuals) will be reset to zero at the beginning of each game; and
12. this process will continue until the 1000 games are completed (see appendix for further detail);

The following diagram, figure 5.3, explains one round of each game (this is repeated for 100,000 rounds and 1000 games):

Figure 5.3 Diagrammatic representation of 1 round of the two-person random interaction model (2PRIM)



The previous section examined the structural and dimensional effects of the three factors in the utility equation. In the simulations that follow, the values chosen for the common good factor (*CGfactor*), selfish factor (*SFfactor*) and cost of competition factor (*Cfactor*) are varied, and results are analysed to see how this changes the levels of selfishness and altruism through evolution. In the first run model the value for these three factors are loosely based on the concept of returns to scale, with *CGfactor* = 1 and *SFfactor* =1, but *Cfactor* = 0.25. These values of *CGfactor*, *SFfactor* and *Cfactor* meet the requirements of the Prisoners' Dilemma framework that is identified by the equation derived earlier:

$$2.(SFfactor - Cfactor) > CGfactor > (SFfactor - 2.Cfactor) \qquad (14)$$

Note that, because of the Prisoners' Dilemma framework, the more selfish individual will always prevail (gain greater utility, fitness and payoff) in an interaction with the less selfish (more altruistic). This was demonstrated in the previous section. However, the point will be to see precisely how this occurs (1) in an evolutionary process, and (2) as the result of changes in the above factors. Key results will be whether equilibrium levels of selfishness and utility are established, what these levels are and what are the levels of utility and total utility attached to them.

At this point some clear distinctions must be reinforced. The levels of selfishness-altruism for any individual are only ever allocated at random in 2PRIM. They are determined at the start of each experiment and do *not* change thereafter. Therefore, when the following speaks in shorthand of changes in the levels of selfishness and altruism through evolution, it *does not* mean that individuals' randomly-allocated trait values of selfishness-altruism change through evolution. *Rather it means changes in the average levels of all individuals-in-the-pool of 1000's randomly-allocated trait values of selfishness-altruism because of the evolutionary survival-exclusion process described at steps 9-12 above*. It is about survival of individuals with higher or lower values of the altruism-selfishness behavioural trait.

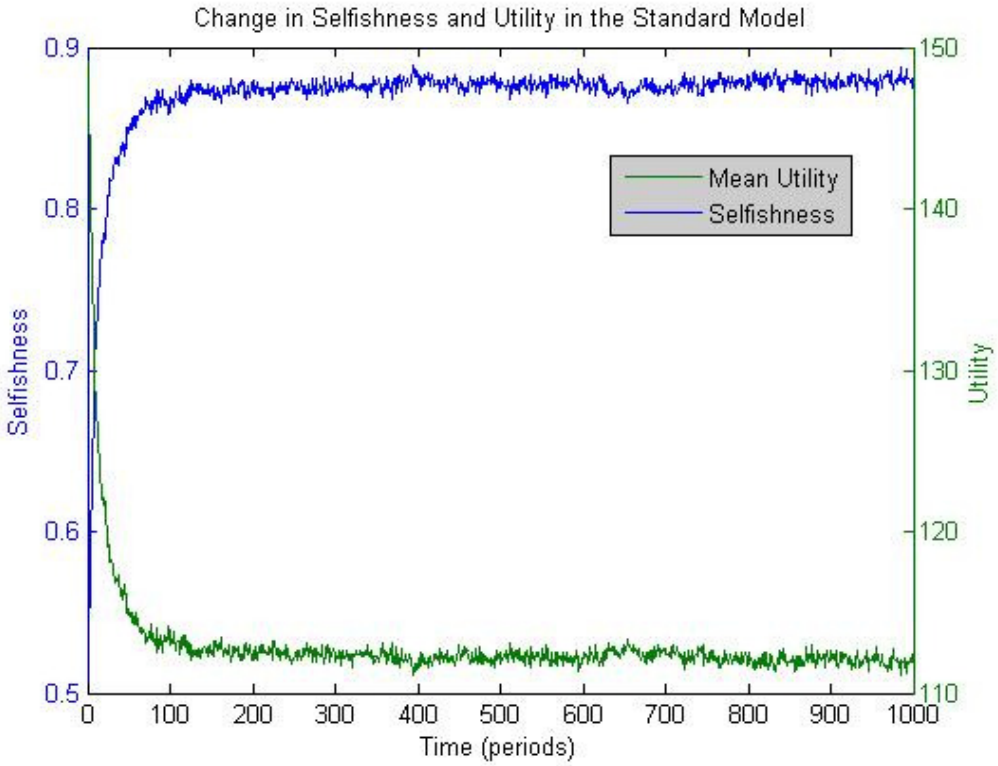
In the standard 2PRIM in figure 5.4 below, the value of the common good factor (*CGfactor*) equals 1.00, selfish factor (*SFfactor*) equals 1.00 and the cost of competition factor (*Cfactor*) equals 0.25. In this version of the model there is no increased return (or public-good characteristic) associated with altruism, but there is a cost incurred when two selfish individuals meet. In the simulation (encompassing 1,000 games with 100,000 rounds in each game, as described above), it was found that the average level of selfishness increased rapidly within the first 100 games, and equilibrium was attained by the 100^{th} game. Within the first 100 games, altruists represent a much higher

proportion of individuals eliminated through the evolutionary process. At equilibrium, the average level of selfishness for the surviving individuals is about 88 per cent (altruism about 12 per cent). However, as figure 5.4 shows, total utility declines as selfishness increases.

In order to assist in better understanding these graphs (provided below), I would like to state that the X-axis shows the number of time periods for which the experiment is undertaken with each time period representing 1 game (i.e. 100,000 two-person random interactions). However, the first Y-axis (left hand side) represents the level of selfishness/altruism in the game and the second Y-axis (right hand side) provides the mean level of utility of the pool in each time period.

An increase in the level of selfishness is consistent with the results provided by Guth and Yaari (1992) in a two-person Prisoners' Dilemma game and Gintis (2000) and Fehr and Gächter (2000) in an N-person game. This occurs because free riders act selfishly to improve their utility. In these authors' games, punishment is added to improve the level of altruism within the game. However, in Gintis (2000), the equilibrium level of selfishness without punishment is consistent with the results in this model. However, the results of 2PRIM are the opposite of those provided by Eldakar et al. (2007), in which altruism starts at around 100 per cent in the first few generations and an equilibrium is attained at the 500^{th} generation, where there is 80 per cent altruism and 20 per cent selfishness.

Figure 5.4 Level of selfishness and utility in the standard 2PRIM



Eldakar et al. (2007) explain these results by saying that the most selfish individuals are removed within the first 20 generations, then punishers are eliminated by the 50th generation due to the cost of punishment that they incur and due to the negative correlation between altruism and punishment, we see only altruistic non-punishers and selfish punishers survive. In Eldakar et al.'s (2007) model altruism increases initially due to the threat of punishment. The standard 2PRIM model, as described in this chapter, does not use punishment. Naturally selfishness succeeds and utility is less than it otherwise would be. When punishment is applied in 2PRIM in chapter 6 we can observe whether the results are consistent with those provided by Eldakar et al. (2007).

In figure 5.4, selfishness will *never* reach 100 per cent and altruistic individuals will always exist as new altruistic individuals enter the group at the end of each game. Recall that 10 per cent of the individuals with the least utility in the pool of 1000 individuals are removed and replaced by new individuals with random levels of the selfishness and altruism trait values.

In summary, selfishness will dominate in this particular ('standard' 2PRIM) evolutionary game. This is despite the fact that the competing selfish individuals suffer due to the (relatively small) cost of competition. Constant returns to altruism (or a *CGfactor* of one) could occur in society where the government safety net is low, like in many developing countries. On the other hand, in many developed countries the government taxes individuals at a higher rate and redistributes these taxes to provide services for all individuals in society. This helps to increase the common good, as people are mandated to contribute to society. Hence it will be useful to compare the results in figure 5.4 with those following an increase in the return on the common pool investment (to a *CGfactor* value greater than 1.00), while holding the *SFfactor* and *Cfactor* constant at 1.00 and 0.25, respectively. What we are looking for specifically is whether an increase in the return to altruism (by increasing the *CGfactor*) will lead to an increase in the number of altruistic individuals who survive in successive games. The questions are whether and at what point the altruists might dominate the game.

Hirshleifer and Rasmusen (1989), Levine (1998), Sethi and Somanathan (2005) and Danielson (2002) suggest that an increase in individual morality is needed to increase the level of co-operation within a repetitive game. Reciprocity must be in some way underlying the cooperative actions of the altruists if a stable equilibrium is to exist. However, in this standard form of the 2PRIM model, which does not include reciprocity and punishment, such a change in morality would have to involve increasing the levels

of altruism (and decreasing selfishness) of individuals in the pool. In other words, it would be outside the random attribute-allocation process of 2PRIM. Nevertheless we can test to see whether, without changing the random altruism-selfishness trait values, we can obtain a similar result (i.e. an increase in the average level of surviving altruism) by changing the *CGfactor* to a value greater than 1.00. That is, can the average level of altruism in the model increase and, possibly, dominate if reciprocity is not required?

While, the standard 2PRIM model meets the conditions of the Prisoners' Dilemma model, these conditions are breached once the CGfactor, SFfactor and Cfactor values are altered as discussed earlier. The intention of this chapter is to explain the boundaries of the 2PRIM model and how the game evolves when these factors are increased from those provided in the standard 2PRIM model. Chapter 6 will then analyse how strong reciprocity can be included in 2PRIM and if it is possible to improve common good using the CGfactor, SFfactor and Cfactor compared to using punishment. We will now analyse the 2PRIM model when the Prisoners' Dilemma conditions no longer hold. Here, as per my discussion with Omar Eldakar, he explained that there are situations where even the most selfish person will have to act altruistically. For example, if two strangers get stuck in a lift and there is a fire in the building. The only way they can get out of the lift is if both work together and exit the lift from the ceiling of the lift. Even the most selfish individual will have to work with the other player to exit this lift. Obviously, a selfish individual will try to get out of the lift first, compared to a more altruistic individual. Similarly, even the most altruistic individual can act selfishly as well. For example, if there is an earthquake, even the most altruistic individual will try to run out of the building (while pushing other people out of the way in panic) to make sure he/she does not get crushed if the building collapses.

So, figures 5.5 to 5.7 illustrate the result with *CGfactor* = 1.00 versus the same interaction with *CGfactor* increased first to 1.50 and then to 2.0. The second and fourth panels of figures 5.6 and 5.7, namely 'Utility, fitness, payoff for each player' and 'Change in player utility, fitness, payoff compared with CGF = 1.00', demonstrate the following:

1. interactions with the greatest levels of total altruism gain the most from increasing *CGfactor*;
2. This is so regardless of the fact that the individuals with the highest selfishness trait value get the highest payoffs in all interactions. We shall see below how this translates into the evolutionary model.

Figure 5.5 2PRIM formulae for utility from interactions for CGfactor = 1.0

Altruism, selfishness traits for each player

		Player 1 A	Player 1 S
Player 2	A	A_1 , A_2	S_1 , A_2
	S	A_1 , S_2	S_1 , S_2

Nominal selfishness, altruism traits for each player

		Player 1 A	Player 1 S
Player 2	A	1.0 , 1.0	1.0 , 1.0
	S	1.0 , 1.0	1.0 , 1.0

Utility, fitness, payoff for each player

		Player 1 A	Player 1 S
Player 2	A	1.00 , 1.00	1.25 , 0.25
	S	0.25 , 1.25	0.50 , 0.50

CGF = 1.00
SF 1.00
CF 0.25

Total utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	2.00	1.50
	S	1.50	1.00

CGF = 1.00
SF 1.00
CF 0.25

Change in player utility, fitness, payoff

		Player 1 A	Player 1 S	CGF = 1.00
Player 2	A	0.00 , 0.00	0.00 , 0.00	SF 1.00
	S	0.00 , 0.00	0.00 , 0.00	CF 0.25

Figure 5.6 2PRIM formulae for utility from interactions for CGfactor = 1.5

Nominal selfishness, altruism traits for each player

		Player 1 A	Player 1 S
Player 2	A	1.0 , 1.0	1.0 , 1.0
	S	1.0 , 1.0	1.0 , 1.0

Utility, fitness, payoff for each player

		Player 1 A	Player 1 S	CGF = 1.50
Player 2	A	1.50 , 1.50	1.50 , 0.50	SF 1.00
	S	0.50 , 1.50	0.50 , 0.50	CF 0.25

Total utility, fitness, payoff

		Player 1 A	Player 1 S	CGF = 1.50
Player 2	A	3.00	2.00	SF 1.00
	S	2.00	1.00	CF 0.25

Change in player utility, fitness, payoff

		Player 1 A	Player 1 S	CGF = 1.50
Player 2	A	0.50 , 0.50	0.25 , 0.25	SF 1.00
	S	0.25 , 0.25	0.00 , 0.00	CF 0.25

Figure 5.7 2PRIM formulae for utility from interactions for CGfactor = 2.0

Nominal selfishness, altruism traits for each player

		Player 1 A	Player 1 S
Player 2	A	1.0 , 1.0	1.0 , 1.0
	S	1.0 , 1.0	1.0 , 1.0

Utility, fitness, payoff for each player

		Player 1 A	Player 1 S	CGF = 2.00
Player 2	A	2.00 , 2.00	1.75 , 0.75	SF 1.00
	S	0.75 , 1.75	0.50 , 0.50	CF 0.25

Total utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	4.00	2.50
	S	2.50	1.00

CGF = 2.00
SF 1.00
CF 0.25

Change in player utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	1.00 , 1.00	0.50 , 0.50
	S	0.50 , 0.50	0.00 , 0.00

CGF = 2.00
SF 1.00
CF 0.25

Figure 5.8 below models evolutionary changes in the levels of average selfishness and utility when the *CGfactor* is increased from 1.00 to 1.50 (*SFfactor* = 1.00; *Cfactor* = 0.25). In figure 5.4 we noticed that the average levels of selfishness increased and utility surviving decreased within the first 100 games. After 100 games a steady state was reached. However, when the *CGfactor* is raised to 1.50, more altruistic individuals obtain higher utility and start surviving in the evolutionary game. However, an increase in the *CGfactor* to 1.5 is not sufficient for the altruistic population to dominate. This repetitive interaction between the return to selfishness (as in the standard 2PRIM) and the increase in altruism from the higher *CGfactor* causes the level of selfishness and altruism to oscillate around 50 per cent. The mean level of utility also oscillates between the levels of 192 and 206.

When the *CGfactor* is raised to 2.00, altruism (i.e. the survival of individuals with higher altruistic trait values) now dominates, with the mean level of the selfish trait value after 100 rounds dropping to 0.2500 (i.e. altruism = 1 – 0.2500 = 0.7500). The mean utility in the game also increases to 325 (as seen in figure 5.9 below). It is found that there is also a near perfect negative correlation of 0.999 between the level of selfishness-altruism surviving and mean utility in this experiment.

Figure 5.8 Level of selfishness and utility in 2PRIM at CGfactor = 1.5

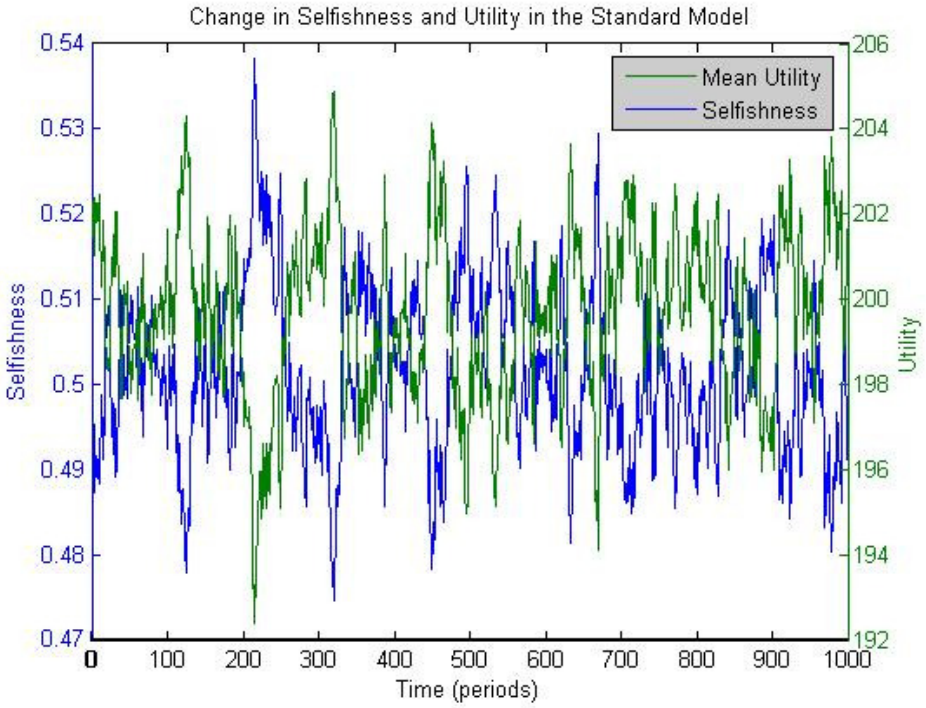
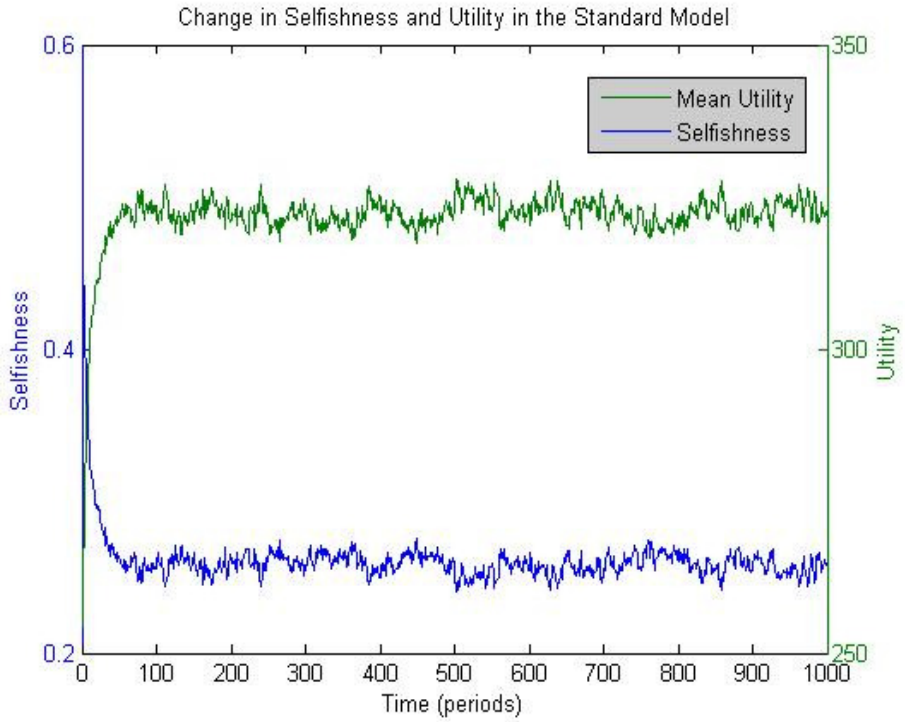


Figure 5.9 Level of selfishness and utility in 2PRIM at CGfactor = 2.0



There are two forces at work. The increase in returns to altruism increases utility for both the altruistic and the selfish. However, the return to the selfish individuals is substantially less as a proportion of total utility. Secondly, through evolution, altruists start to dominate with more selfish individuals finding themselves out of the game. As a result altruists are more likely to meet like minded individuals to the benefit of both (in terms of utility gained). At a CGfactor greater than 1.5 (i.e. 1.6 or 1.7) the level of altruism clearly increases and is consistently sustained above 50 per cent.

Such an instance can be seen in human societies, where people work together in companies, groups, clubs to increase their own utility and that of others. While, evolution occurs, a sustainable equilibrium is attained where people work together and some level of common good is attained. There are also conditions where the level of common good is higher than normal, for example, in the case where people donate funds for a common cause like finding a cure for cancer. In such a circumstance, altruism will survive more than selfishness as individuals who act selfishly will not be liked by other people in society. While, we have analysed how an increase in the CGfactor affects the 2PRIM results, it is important to look at the impact on these results provided by the 2PRIM model will be when SFfactor and Cfactor are changed. Will an increase in these two factors reduce the mean utility and level of selfishness within the evolutionary game?

5.4 Analysing the selfish factor and Adam Smith's impartial spectator (guilt)

As discussed earlier, any increase in the selfish factor (SFfactor) violates the conditions of the Prisoners' Dilemma framework. However, there are numerous conditions in real life where such violations of the Prisoners' Dilemma framework occur. For example, in

the ATP US open championship, players meet through a random draw and compete against each other. Each individual wants to win the grand slam and be ranked the number one tennis player in the world. Therefore, each player will act selfishly in order to have the best chance of winning the tournament.

As a result, while the level of selfishness in the standard 2PRIM model reaches 88 percent in the experiment at the end of 100 iterations (as noted earlier), an increase in the SFfactor has an impact on the level of selfishness within the game and as expected, an increase in the SFfactor further increases the dominance of selfish individuals. When the SFfactor is increased from 1.0 to 1.5 it is found that the level of selfishness increases above 0.9 and this can be seen in figures 5.10 and 5.12 (below). However, this is only slightly higher than that of the standard model (where it is around 0.88). Further increases appear to have little effect on the mean level of selfishness in the game. This occurs, in part, due to evolutionary process adopted in 2PRIM. In addition to this, the cost of competition reduces the mean utility for the most selfish individuals who find themselves removed from the game.

Figure 5.10 2PRIM formulae for utility from interactions for SFfactor = 1.5

Nominal selfishness, altruism traits for each player

		Player 1 A	Player 1 S
Player 2	A	1.0 , 1.0	1.0 , 1.0
	S	1.0 , 1.0	1.0 , 1.0

Utility, fitness, payoff for each player

		Player 1 A	Player 1 S
Player 2	A	1.00 , 1.00	1.75 , 0.25
	S	0.25 , 1.75	1.00 , 1.00

CGF = 1.00
SF 1.50
CF 0.25

Total utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	2.00	2.00
	S	2.00	2.00

CGF = 1.00
SF 1.50
CF 0.25

Change in player utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	0.00 , 0.00	0.50 , 0.00
	S	0.00 , 0.50	0.50 , 0.50

CGF = 1.00
SF 1.50
CF 0.25

Figure 5.11 2PRIM formulae for utility from interactions for SFfactor = 2.0

Nominal selfishness, altruism traits for each player

		Player 1 A	Player 1 S
Player 2	A	1.0 , 1.0	1.0 , 1.0
	S	1.0 , 1.0	1.0 , 1.0

Utility, fitness, payoff for each player

		Player 1 A	Player 1 S
Player 2	A	1.00 , 1.00	2.25 , 0.25
	S	0.25 , 2.25	1.50 , 1.50

CGF = 1.00
SF 2.00
CF 0.25

Total utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	2.00	2.50
	S	2.50	3.00

CGF = 1.00
SF 2.00
CF 0.25

Change in player utility, fitness, payoff

		Player 1 A	Player 1 S
Player 2	A	0.00 , 0.00	1.00 , 0.00
	S	0.00 , 1.00	1.00 , 1.00

CGF = 1.00
SF 2.00
CF 0.25

In contrast, Eldakar et al. (2007) have explained that selfishness is not the dominant equilibrium and that altruism dominates when there is a threat of punishment. In 2PRIM as described in this chapter, punishment is not used (though it will be applied to this model in the following chapter). Therefore, selfishness is the stable equilibrium in this game as seen in figure 5.4 and when the SFfactor is increased, it just reinforces the dominance of the selfish equilibrium.

Adding to this discussion, Becker (1981) states that altruism exists in interactions with related individuals (e.g. family) and selfishness dominates in interactions with unrelated individuals (e.g. business transactions). The results from 2PRIM align with Becker (1981) in the standard model and when the SFfactor is increased above the value of 1.00. However, it contradicts Becker (1981) when the CGfactor and Cfactor are increased. This happens as altruism increases as the return on the common pool and the cost of competition reduce the value in being selfish, resultantly a stable altruistic equilibrium is attained in that instance.

Bowles (2008) also says that selfish actions are important as they drive an individual's desire to progress, however these selfish actions should align to the altruistic frameworks in which society operates. The results of 2PRIM are consistent with Bowles (2008) where selfish actions underlie the standard 2PRIM and selfishness increases with an increase in the SFfactor. However, an altruistic equilibrium is attained when individuals co-operate due to an increase in CGfactor that relates to socially efficient outcomes.

Figure 5.12 Level of selfishness and utility in 2PRIM at SFfactor = 1.5

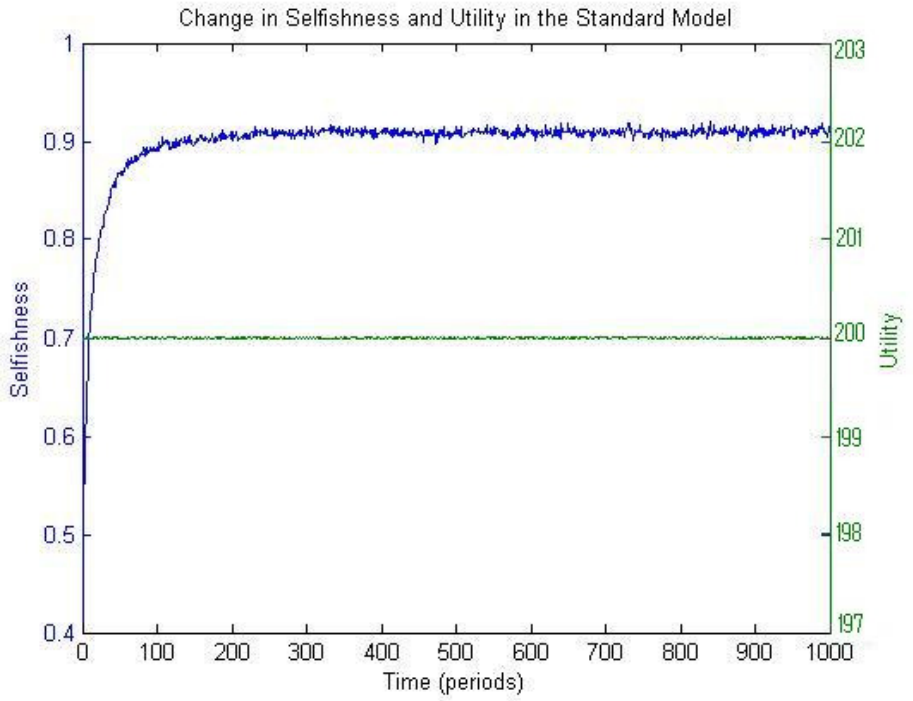
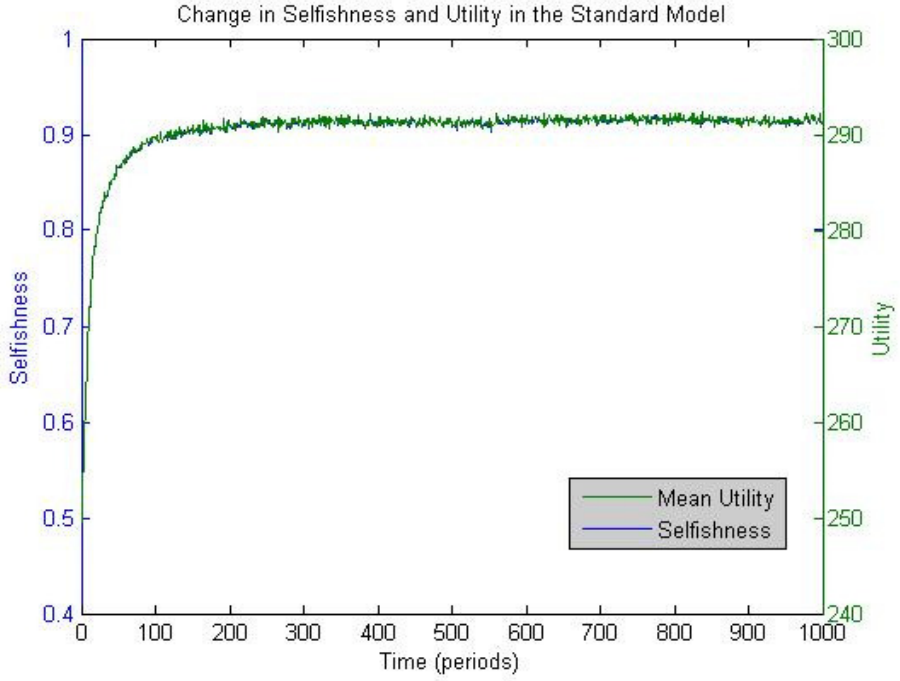


Figure 5.13 Level of selfishness and utility in 2PRIM at SFfactor = 2.0



Further, there is no correlation between the level of selfishness and mean utility in figure 5.12. This occurs due to the fact that while in the standard 2PRIM the level of utility decreases rapidly within the first 100 games (see figure 5.4), in this case though the higher return on the selfish investment (SFfactor = 1.5) equals this drop in utility. This results in the mean utility staying constant at 200 from games 1 to 1000. This is a surprising result as the Selfish Factor (SFfactor) has now started reinforcing the level of selfishness within the game. So, selfish individuals get a higher return for being more selfish. Therefore, even the highly selfish individuals (with their level of selfishness greater than 0.90) are now getting higher utility through the higher SFfactor and the cost of competition (Cfactor) is still the same (i.e. Cfactor = 0.25; as in the standard model). As a result, the mean utility of these selfish individuals' increases and a greater number of selfish individuals survive. This can be further seen in figures 5.10 and 5.13 where the SFfactor = 2.0. In this case an increase in the SFfactor directly increases the level of utility of the selfish individuals within the game. While the SFfactor has an impact on the mean utility and the level of selfishness within the game, a change in the Cfactor can have a significant impact on the evolution of selfish individuals as it increases the cost of acting selfishly. Let's analyse how this increase in the Cfactor affects the results from the 2PRIM model.

5.5 Changes in the cost of competition and level of selfishness

It is clear from the previous results that destructive competition can have a significant impact on the world around us. Competition is a useful factor in life that motivates people and pushes people to deliver better results than that could be obtained by them alone. However, competition starts to become destructive when it causes more harm

than good, for example, where a competitor wants to win the game at any cost, regardless of the harm that they themselves or others may face. Such competition is unproductive and the cost of competition factor (Cfactor) then contributes negatively to the utility of selfish individuals in games, in instances where they meet other selfish individuals. At times, when these selfish individuals meet altruistic individuals the cost of competition is low and the disutility for both the players is lower. An increase in the Cfactor violates the conditions of the Prisoners' Dilemma model as discussed earlier. However, there are real world situations where players will be highly competitive as to cause destructive competition. For example, consider a Formula 1 or NASCAR race. Drivers want to win the World Championship and often take risks to overtake the car ahead of them. If a driver is too competitive and tries to push his way in front of the car ahead, he could instead end up causing a car crash. In this case, neither he nor the car in front ends up winning the race.

Seminal work by Reeve (1906) describes the 'cost of competition' as the cost to business enterprise of goods and work. The 2PRIM uses the cost of competition factor (Cfactor) consistent with the definition proposed by Reeve (1906). Results provided by Eldakar et al. (2007) and Eldakar and Wilson (2008) also align with the concept of a cost of competition, where selfish punishers compete with other selfish individuals and the cost of competition is effectively applied by eliminating the more selfish individual from the game. Guth and Yaari (1992), Fehr and Schmidt (1999) and Hirshleifer and Rasmusen (1989) also support the idea that there is a cost for destructive competition.

When, the Cfactor increases above 0.25, it is seen that selfishness starts to fall. As, the cost of competition increases, the utility of selfish individuals decreases rapidly and the level of selfishness decreases in the game as the highly selfish individuals start getting eliminated. An increase in Cfactor also has a negative impact on the utility of

the altruists, when these altruists meet selfish individuals. The only instance where Cfactor will not have an impact on the utility of the two players in 2PRIM is when these two players are 100 per cent altruistic (as Cfactor only provides for negative utility for selfish behaviour). This can be seen in figures 5.14 to 5.17 that provide three dimensional mesh graphs with these snapshots taken at the end of each experiment (i.e. 1000 games). In figure 5.14, where the cost of competition (Cfactor) is equal to 0.00, we see that the level of selfishness is high and selfishness starts to increase at a SFfactor = 0.2. This happens as selfish individuals gain higher utility even at lower SFfactor levels.

Figure 5.14 Level of selfishness in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.00

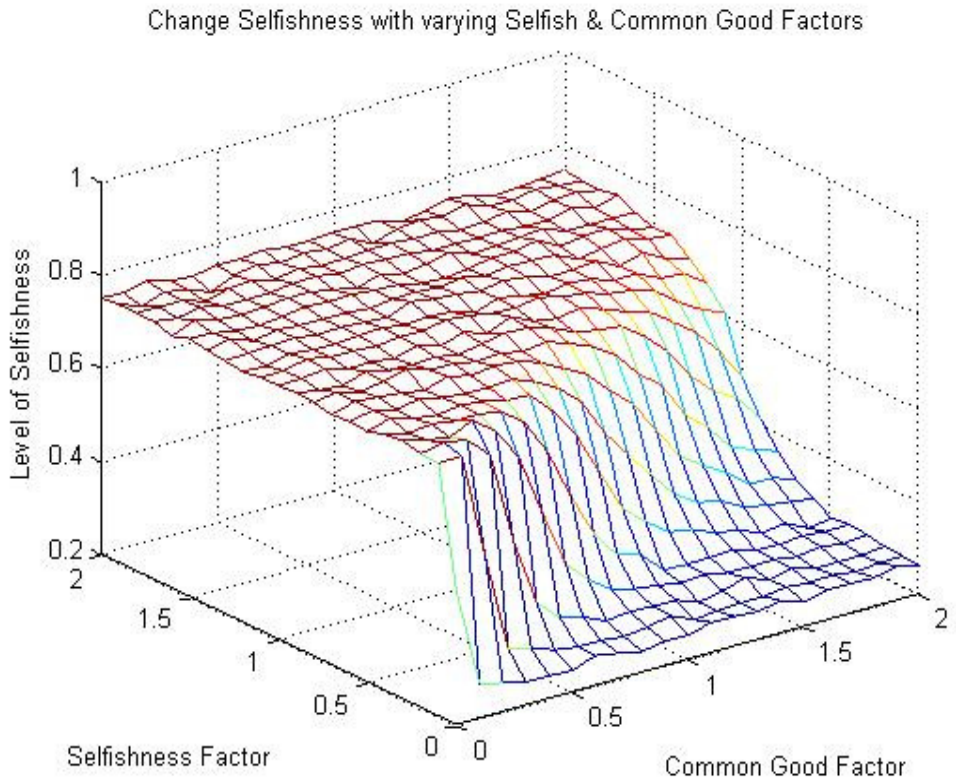
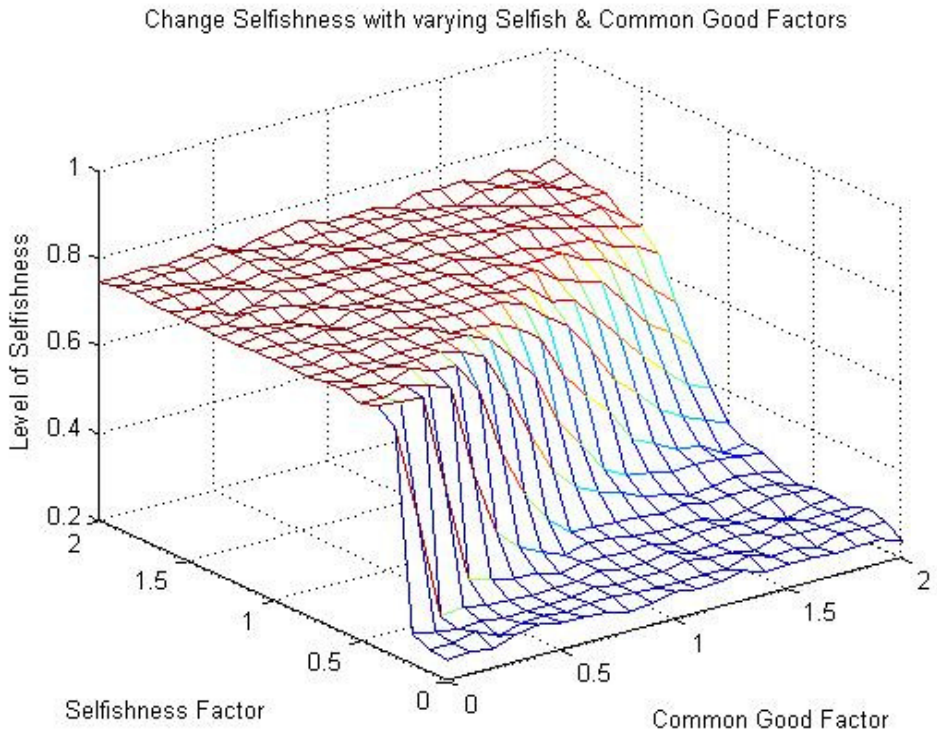


Figure 5.15 Level of selfishness in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.25



The level of selfishness also increases quickly once the SFfactor increases above 0.2 with the level of selfishness reaching the value of 0.78 (at SFfactor = 2.0). In figure 5.14, we see that selfish individuals dominate the system when the cost of competition is nonexistent. In a society where there is no disutility for destructive competition, it would make sense for people to be selfish. However, as we cannot do everything by ourselves we need to work in groups (e.g. companies, unions, clubs etc) to accomplish and tackle the more difficult and complex tasks in our environment. This forces us to co-operate and work in a group as discussed earlier. In such instances, it will be fruitless to have highly selfish people involved in a group as they will usually try to take advantage of other individuals in the group and the group will either remove this person or force the selfish person to co-operate. So, you can see while selfish individuals may do well in the case where Cfactor = 0.00, this is not a realistic case in society.

As, Cfactor is raised to 0.25 (which is the base case), selfishness decreases in the system. In figure 5.15, we see that at Cfactor = 0.25, the level of selfishness is lower and the selfishness trait only dominates after the SFfactor increases above SFfactor = 0.50. On the contrary, when the Cfactor = 0.00 (in figure 5.14), selfishness increases earlier (near SFfactor = 0.2) compared to this case (when Cfactor = 0.25), where the level of selfishness increases at SFfactor = 0.50. This occurs as the most selfish individuals now obtain lower utility due to the higher cost of competition. Figure 5.15 shows that the majority of the individuals tend to be altruistic at higher CGfactor levels compared to what we see in figure 5.14.

The level of selfishness reduces much further as the Cfactor is doubled to 0.50 (as seen in figure 5.16). At Cfactor = 0.50, the level of selfishness in the game has substantially decreased and selfishness only increases once the SFfactor is greater than 0.5. This happens as the high cost of competition prevents selfishness from increasing at lower SFfactor levels. At higher CGfactor levels in figure 5.16, the level of altruism also increases and when CGfactor = 2.00, selfishness only arises when SFfactor is greater than 1.8. This graph (figure 5.16) shows that an increase in the cost of competition factor (Cfactor) to the level of 0.5 results in the level of selfishness dropping substantially, as the more selfish individuals get a higher level of disutility from the higher cost of competition.

Figure 5.16 Level of utility in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.50

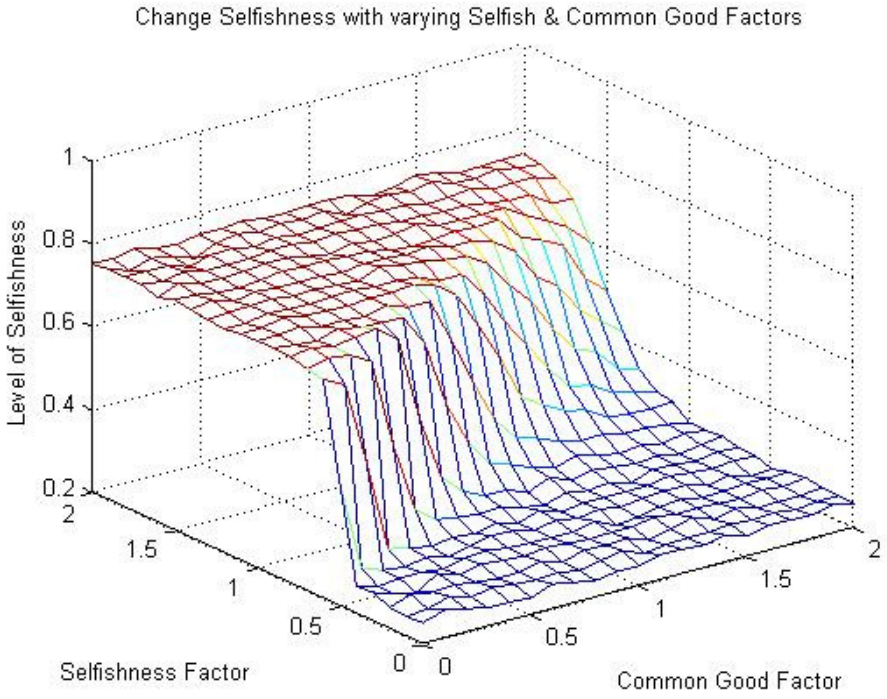


Figure 5.17 Level of selfishness in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.75

In society, we would usually see this in the case of an oligopolistic market, where the participants in the market will not be selfish because the other participants have similar market power and the cost of competition is high. Therefore, participants in oligopolistic markets (at Cfactor = 0.50) will tend to co-operate more with their competitors compared to when there is no cost of competition (i.e. Cfactor = 0.00; as in figure 5.14). At Cfactor = 0.00, the majority of the individuals are selfish (in figure 5.14), while at Cfactor = 0.50, the minority of the individuals are selfish and the majority are altruistic (in figure 5.16). The level of selfishness decreases at higher Cfactor levels because the destructive cost of competition between selfish individuals substantially increases and this higher Cfactor forces the more selfish individuals to obtain lower utility that results in them getting eliminated in subsequent games.

We have discussed how the change in Cfactor can affect the *level of selfishness* within 2PRIM. Now, it will also be useful to understand how this affects the *change in mean utility* within these games.

5.6 Cost of competition and mean utility

When the Cost of Competition (Cfactor) equals 0.00, the mean utility is rather high at 400 because individuals gain from both an increased return from the SFfactor (return from selfish investment) and CGfactor (common pool return) as seen in figure 5.18. Though, the mean utility falls quickly when the Cfactor increases. As seen in figure 5.19, when the Cfactor increases to 0.25, the highest level of mean utility decreases from 400 to 300. This happens as higher cost of competition reduces utility from increased selfishness. As the level of selfishness increases (see figure 5.15) we notice that the mean utility drops. This would be expected as we have discussed that an

increase in destructive competition will reduce the utility for selfish individuals. As the decrease in utility for selfish individuals relates to an increasing level of selfishness, we notice in figures 5.14 – 5.17 that a greater number of selfish individuals are eliminated in subsequent games.

As, higher levels of Cfactor (above 0.25) are a violation of the Prisoners' Dilemma framework, nonetheless such situations do occur in reality, for example, in oligopolistic markets (like the Australian banking or telecommunication sector), competitors will reduce the price of their good if one of the other firms drops its prices. Effectively, this will cause disutility to all firms in the oligopolistic market. The cost of competition will be high in any situation where there is strong competition and in instances where competitors have a possibility of challenging other competitors by causing disutility.

While, increasing Cfactor values cause a disutility for selfish behaviour, we also notice that there is a slight depression in the utility graph in figure 5.18 that becomes deeper as the Cfactor increases (see figures 5.19 and 5.20). This depression is caused by the increase in the level of selfishness as the SFfactor increases (see figures 5.14 to 5.17) and coincides with an increase in the Cfactor, as higher selfishness results in lower mean utility. You will also notice that as the Cfactor increases, the mean utility decreases from 400 (in figure 5.18) to 200 (in figure 5.21) due to the disutility caused by the cost of competition related to the level of selfishness, so the more selfish the individual the more likely that they will have a lower utility.

Figure 5.18 Level of utility in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.00

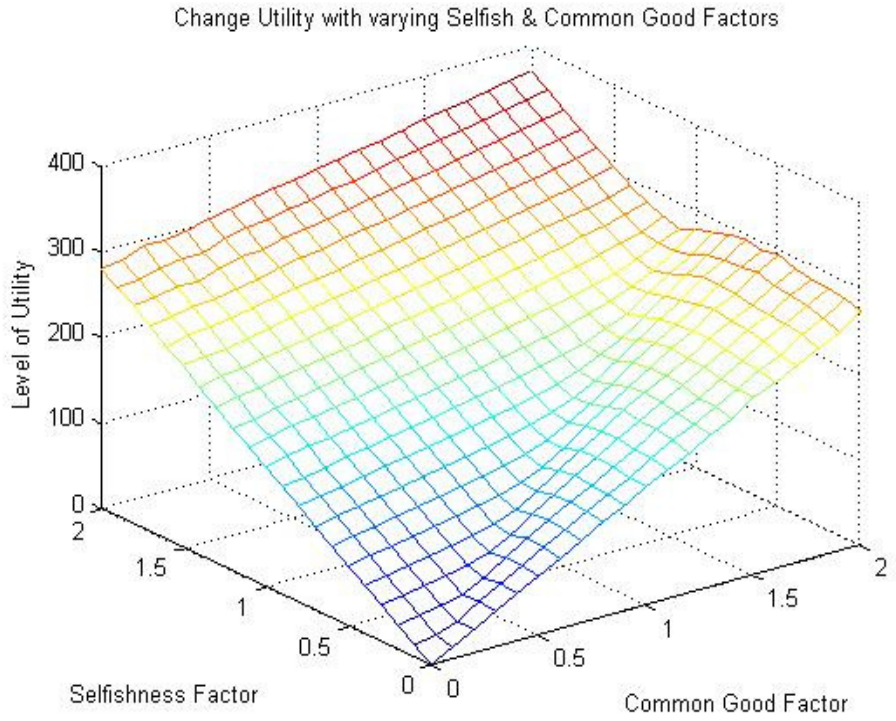


Figure 5.19 Level of utility in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.25

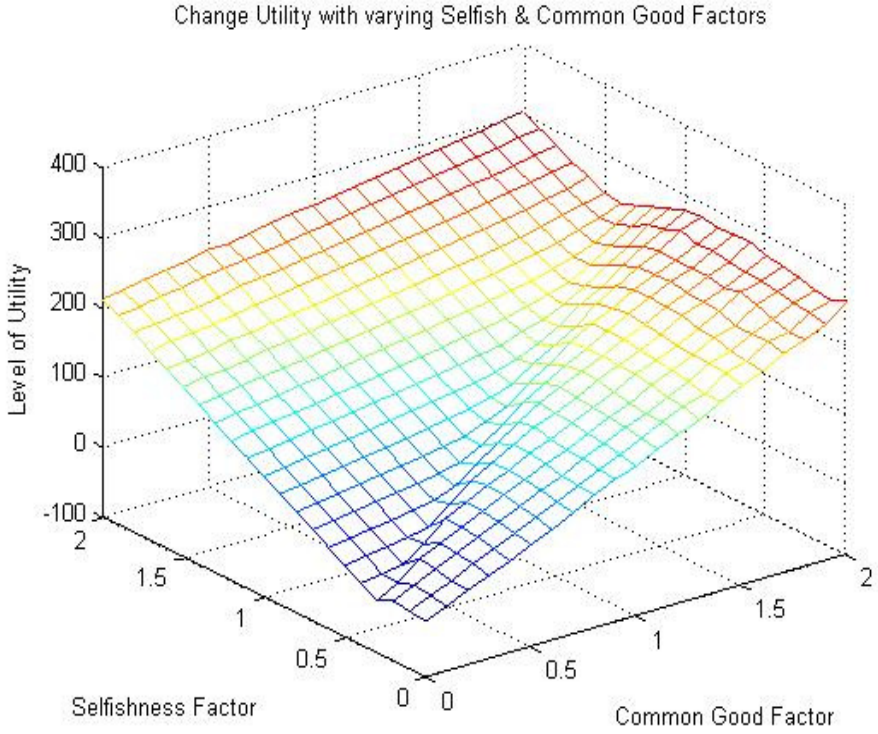


Figure 5.20 Level of utility in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.50

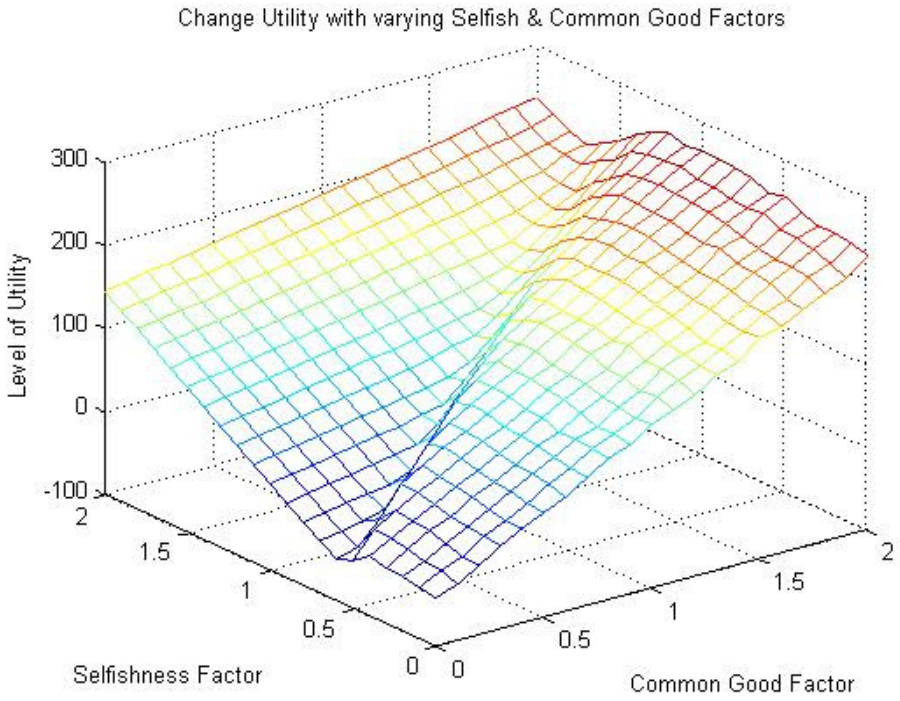
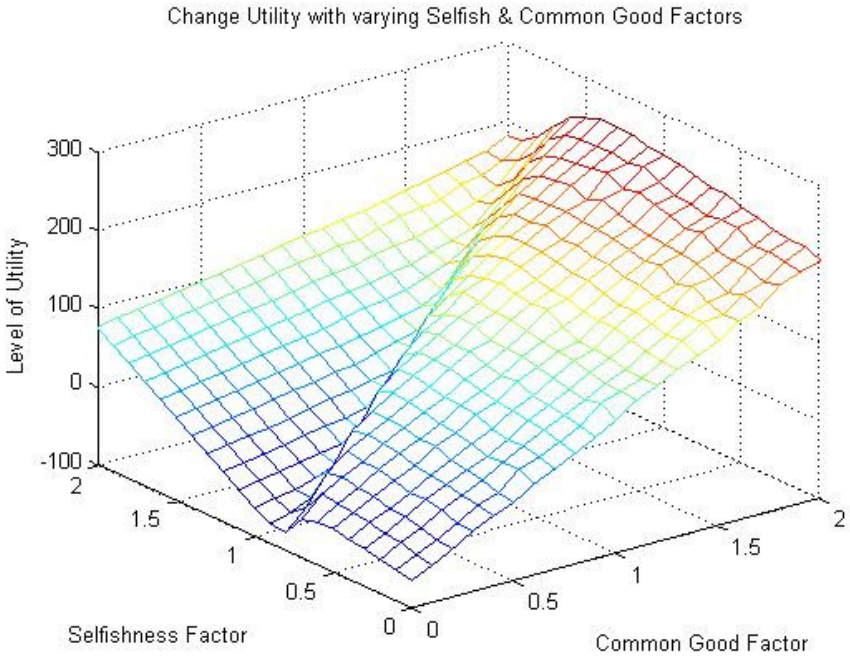


Figure 5.21 Level of utility in 2PRIM with different levels of the selfish (SFfactor) and CGfactor, when Cfactor = 0.75



The maximum level of utility decreases surprisingly as the altruistic individuals lose utility due to the cost of competition; primarily due to the fact that not all altruistic individuals are 100 per cent altruistic. These altruistic individuals have some selfishness themselves and they receive disutility for how selfish they are due to the random interaction they have with a selfish individual. Collectively, this causes the mean utility of the system to reduce.

Also, we notice that utility drops off as the SFfactor increases and the depression becomes more significant as we move from figure 5.18 to 5.21. This depression occurs as a phase transition takes place where the level of selfishness increases rapidly (see figures 5.14 – 5.17).

5.7 Conclusion

Bowles and Gintis (2000) coined the concept called strong reciprocity where individuals would punish others in order to increase the altruism within the game; this is otherwise called second-order altruism. Guth and Yaari (1992), Fehr and Gächter (2000), Boyd et al. (2003) and others have discussed that kin selection; genetic co-evolution and strong reciprocity are some of the strategies that increase altruism within N-person and two-person games. Eldakar et al. (2007) and Eldakar and Wilson (2008) have stated that only selfish individuals are successful in punishing other selfish individuals in order to improve their utility. Therefore, they classify this as second-order altruism rather than second-order selfishness. This chapter develops the two-person random interaction model (2PRIM) which is based on the two-person Prisoners' Dilemma model, where two individuals meet each other at random and their utility is calculated based on their level of altruism/selfishness (i.e. A-S trait value). Results

show that selfishness increases rapidly and these results are similar to Guth and Yaari (1992), Eldakar et al. (2007) and Eldakar and Wilson (2008).

Further, three additional factors labelled the return on common good (CGfactor), return on selfishness investment (SFfactor) and cost of competition factor (Cfactor) are discussed. Based on the standard 2PRIM model, the level of selfishness increases to 88 per cent and reaches a stable equilibrium after 100 games. However, the Prisoners' Dilemma framework on which the standard 2PRIM model is based on is violated when the values of the CGfactor, SFfactor and Cfactor are increased to analyse other real world situations. However, this provides us an interesting insight into the evolutionary boundaries of the 2PRIM model and explains results of real world 2 person random interactions outside the boundaries of the Prisoners' Dilemma framework, which is one of the key contributions of this 2PRIM model.

Results show that if the return on common (public) good (CGfactor) is marginally increased in 2PRIM, the level of altruism increases, but oscillates around 50 per cent. This occurs due to the fact that the increase in the return on common (public) good is insufficient for altruism to take over. So, there is competition between selfishness and altruism to overtake each other. However, when the return on the common (public) good is increased further, then altruism and utility significantly increases and is nearly double of that found in the base case. This occurs due to the fact that it now pays for individuals to invest in the common (public) good compared to the selfish investment. So, altruistic individuals who invest in the common good end up getting a higher return, compared to the selfish individuals who mainly invest in the selfish investment.

Similarly, if the return on the selfish investment (SFfactor) is increased, the level of selfishness barely increases from 88 per cent in the base case to 90 per cent in this situation. This happens as altruists are not completely eliminated from the game, as new

altruists enter at the end of each game, where 10 per cent of the individuals with the least utility are eliminated from the game and a new 10 per cent are selected from the pool of N individuals from where the initial population was obtained. Also, some selfish individuals are eliminated at the end of each game due to the cost of competition. However, mean utility does not increase as much as it does when the return on common good is increased. This happens due to the fact that an increase in the selfish investment only increases the return for the selfish individuals compared to an increase in the common good factor that positively affects the utility of both the selfish and altruistic individuals.

However, the highly selfish individuals suffer disutility from the cost of competition. Therefore, the mean utility is lower when the return on selfish investment is higher compared to a higher return on common good. When the return on selfish investment is increased further, it is noticed that selfishness does not increase any further, but the utility is identical to the level of selfishness, as the utility is derived in this game primarily from being selfish. There are still some altruists remaining in this game due to the new altruists entering at the end of each game and due to highly selfish individuals being eliminated due to the cost of competition.

The cost of competition is the third factor in 2PRIM that affects the level of selfishness and utility within the game. When there is no cost of competition, the selfish individuals succeed and obtain higher utility, though still some altruists remain due to the new altruists joining at the end of each game. However, when the cost of competition is increased, it substantially decreases the level of selfishness as the selfish individuals now start obtaining lower utility and start to get eliminated in subsequent games. Additionally, an increase in the cost of competition reduces the utility of the selfish individuals, thus reducing the mean utility in the game.

Finally, we realise that the return on common good, selfish investment and cost of competition have a significant impact on the level of selfishness and mean utility within 2PRIM. In the next chapter, this thesis will examine what affect strong reciprocity has on the level of selfishness and utility within 2PRIM.

CHAPTER SIX

Strong reciprocity in two-person random interactions

6.1 Introduction

The basic form of the two-person random interaction model (2PRIM) was introduced, discussed and tested in the previous chapter. This model was built on the conceptual development in chapters 2-4. In chapter 5, the model was used to understand how different returns to common good, selfishness and cost of competition affected the evolving levels of selfishness and utility in two-person random interactions. Building on the model in chapter 5, strong reciprocity will be incorporated in 2PRIM in this chapter to understand if punishment has an impact on player survival in this model, again in an evolutionary sense. The question is will punishment impact on the survival of selfish individuals? Also, we need to understand if strong reciprocity has a greater impact on the level of selfishness and utility compared to the return on common (public) good (CGfactor), selfish investment (SFfactor) or the cost of competition (Cfactor).

As discussed in chapter 4, punishment is important to reinforce altruism within games and society. Strong reciprocity is applied in an original way in this model by temporarily increasing the level of selfishness of the less selfish individual in 2PRIM in phase two of each round. Recall from the basic form in chapter 5 that each party to the random one-shot two-person interaction acted and (implicitly) responded according to his/her randomly allocated behavioural altruism. What is used here is that the response phase of the one-shot interaction allows the more altruistic player to respond, as it were 'out of character'. By responding more selfishly, the less selfish player (more altruistic) affects a form of punishment. Results show that when punishment is incorporated in

2PRIM, the surviving average level of selfishness marginally decreases. However, there is little effect in raising utility.

Note that, consistently with the models of strong reciprocity discussed in chapter 4, punishing entails a cost to the punisher as well as the punished. The model is modified to account for this effect. When the cost of punishment is increased, it reduces the utility of altruists and the number of altruists who survive in subsequent games. It is seen that strong reciprocity only causes the level of selfishness and utility to change temporarily, while the return on common (public) good, selfish investment or the cost of competition seem to have a lasting effect, as the pay-offs for the selfish and altruistic individuals change permanently.

Strong reciprocity, according to Bowles and Gintis (2000), is the punishment that is conferred on selfish individuals who disregard social norms in society. Strong reciprocity is a significant factor in human relationships in social contexts (Fehr and Gächter 2000), and it would be important to model punishment in 2PRIM. Gintis (2000c), Kahana (2005) and other researchers have shown that strong reciprocity can also occur in one-off games. Such one-off games can have real-world parallels, for example, if someone is going to work during rush hour when the train is full they might push themselves into the train. It is possible that someone in the train or on the platform will reciprocate in kind. This chapter tries to understand how strong reciprocity affects the evolution of selfish or altruistic behaviours in these two-person random interaction situations.

The next section of this chapter will briefly review some of the relevant literature. The third section will then set out the features of 2PRIM with strong reciprocity. The fourth section will analyse the results obtained from the 2-PRIM with strong reciprocity model and the final section will conclude by summarising the man conclusions of this

chapter. Note that this chapter will be less descriptive than was chapter 5. It will use the three-dimensional mesh graphs in section 6.4 and draw conclusions from there more summarily in order to reduce repetition.

6.2 Strong reciprocity and the 2PRIM model

It will be necessary before introducing strong reciprocity into 2PRIM to look at the different approaches of key authors to the question of punishment (see sections 3.3, 3.4, 4.5 and 4.6). Strong reciprocity is a concept developed by Bowles and Gintis (2000) to describe second–order altruistic behaviour, when altruistic individuals will punish selfish individuals in order to increase co-operation. Human co-operation seems to be an evolutionary puzzle, given the problem of free-riding. Bowles and Gintis (2000) propose a way by which we might solve this puzzle. Their view is that strong reciprocity, through which people voluntarily participate in expensive cooperation by punishing non-co-operators, can exist in groups in which individuals are not even related, and this behaviour cannot be explained by theories of kin selection, reciprocal altruism, costly signalling and indirect reciprocity.

In contrast, Burnham and Johnson (2005) believe that the 'strong reciprocity' seen in experiments studying one-shot interactions between unrelated individuals is not a newly documented concept. Rather it is a maladaptive form of reciprocal altruism and has evolved by individual selection. They say that group selection could play a role, but is neither necessary nor sufficient to explain co-operative behaviour in humans. Contemporary behavioural theories regarding reciprocity have been provided by Cosmides and Tooby (1992), Dawkins (1976), Maynard Smith (1982), Williams (1966) and Wilson (2000), which state that non-kin relations can be modelled using self-

interested actors. Since then substantial behavioural theory research has been done to understand strong reciprocity resulting from non kinship, family and sexual relations. See chapter 4 for a more comprehensive discussion of this issue.

Alexander (2005) states that the basis for strong reciprocity is more cultural than biological in nature. In contrast, Bowles and Gintis (2002b) explain that this occurs through the evolution of genetically transmitted behaviours that relate to individual growth. Culturally transmitted behaviours are passed on through group-level activities and social norms. Still other models explain how cultural factors like resource-sharing are critical to the evolution of genetically transmitted altruism through natural selection. Odling-Smee, Laland and Feldman (2003) and Bowles (2000) believe that it may be helpful to represent human cultures and institutional structures in a way that affects this genetic evolution by supporting the creation of a particularly helpful environment.

Laboratory experiments (Fehr and Gächter 2002, Ostrom et al. 1994) and field data (Boehm et al. 1993) have shown that individuals punish non co-operative behaviour even in one-shot interactions. Even though such altruistic punishment increases co-operation in a group, still it creates a dilemma, as existing models suggest that altruistic co-operation between non-genetically related individuals is evolutionarily stable only in small groups. So punishment in one-shot experiments leads to predictions that people will not incur costs to punish other individuals to provide a benefit to a larger group of non-genetically related individuals. See chapter 4 to review the discussion on laboratory experiments related to strong reciprocity. Boyd et al. (2003) explain how punishment can facilitate altruism within a large population in one-off interactions. Chapter 4 discusses this significant paper.

Masclet and Villeval (2006) analyse why costly punishment would occur in non-genetically related groups and investigate if inequality aversion and negative emotions

have a role in determining if punishment will be applied in such a scenario. Their results show that as the level of inequality increases it causes the level of punishment to increase, which leads to a decrease in the level of fitness inequality between the individuals within subsequent games.

Fehr and Gächter (2000) state that free-riding causes negative emotions that could result in punishment being applied even if it is costly. This punishment causes selfishness to reduce and co-operation to increase in the group. The greater the degree of free riding, more the free riders are punished. Bowles and Gintis (2002, 2004) argue that it is more than self-interest that requires humans to punish free riders. Instead, they state that humans will altruistically punish free riders in order to increase co-operation within the group even if they do not directly benefit from it. Bowles and Gintis (2004) suggest that people can have a predisposition to punish others at a cost to themselves and this punishment can help sustain high levels of co-operation. This idea is also supported by laboratory experiments undertaken by Price et al. (2002). Similarly, a laboratory experiment conducted by Burnham and Hare (2005) shows that greater co-operation is developed within the group when every individual's actions in a group are being monitored compared to the group where the individual's actions are not being monitored. McCabe et al. (1996) also demonstrate that reciprocity exists in one-off games. They find that some people are prone to co-operative behaviour as they co-operate in single play games as if they are in a repetitive series of different games. McCabe et al. (1996) also support the argument that reciprocity in repetitive games increases co-operation.

Bowles and Gintis (2000), Fehr and Gächter (2000), Boyd et al. (2003) and others have discussed strong reciprocity as a second-order altruistic trait or second-order altruism. However, Eldakar et al. (2007) have provided an alternative model showing

strong reciprocity as a form of second-order altruism. Strong reciprocators in the model are selfish individuals who eliminate other selfish individuals to reduce the number of selfish individuals in subsequent games in order to increase their payoffs in those games. A strong reciprocator in Eldakar et al. (2007) and Eldakar and Wilson's (2008) view is a selfish individual who is both first-order selfish and second-order selfish.

The examples that Eldakar et al. (2007, p. 204) and Eldakar and Wilson (2008, p. 6985) used to argue that their particular model of strong reciprocity might have real-world parallels are:

> Considerable evidence for altruism maintained by competition among selfish individuals exists for nonhuman species, from insects to vertebrates. Wenseleers et al. describe a 'corrupt policing' strategy in tree wasps Dolichovespula sylvestris, where workers that police other workers lay their own eggs (Wenseleers et al., 2005). Scrub jays that tend to steal caches from other scrub jays are also more defensive of their own caches (Emery and Clayton, 2001). In addition to our empirical study on humans that inspired our simulation model (Eldakar et al., 2006), the history of medieval knights provides a potential historical example of selfish punishment. Much as the knights of old are revered in mythology and popular culture, the first Castellans are better described as selfish thugs who fought among themselves to exploit the defenceless, and therefore altruistic, peasants (Bisson, 1994). As Pope Gregory VII put it during the 11th century (quoted in Bisson, 1994, p. 42), 'Who does not know that kings and princes derive their origin from men ignorant of God who raised themselves above their fellows by pride, plunder, treachery, and murder?'

The above review shows that there are a number of views in relation to strong reciprocity, where strong reciprocators either punish in order to increase altruism in the group or to improve their future payoffs. In general, strong reciprocators are punishing

others to increase altruism (Bowles and Gintis 2000) or due to their own self-interest (Eldakar et al. 2007). The next section applies punishment to 2PRIM in order to understand what affect punishment has in the way 2-PRIM models two-person random interactions. Will it change the average level of selfishness (altruism) or average utility in an evolutionary context?

6.3 Two-person random interaction model (2PRIM) with strong reciprocity

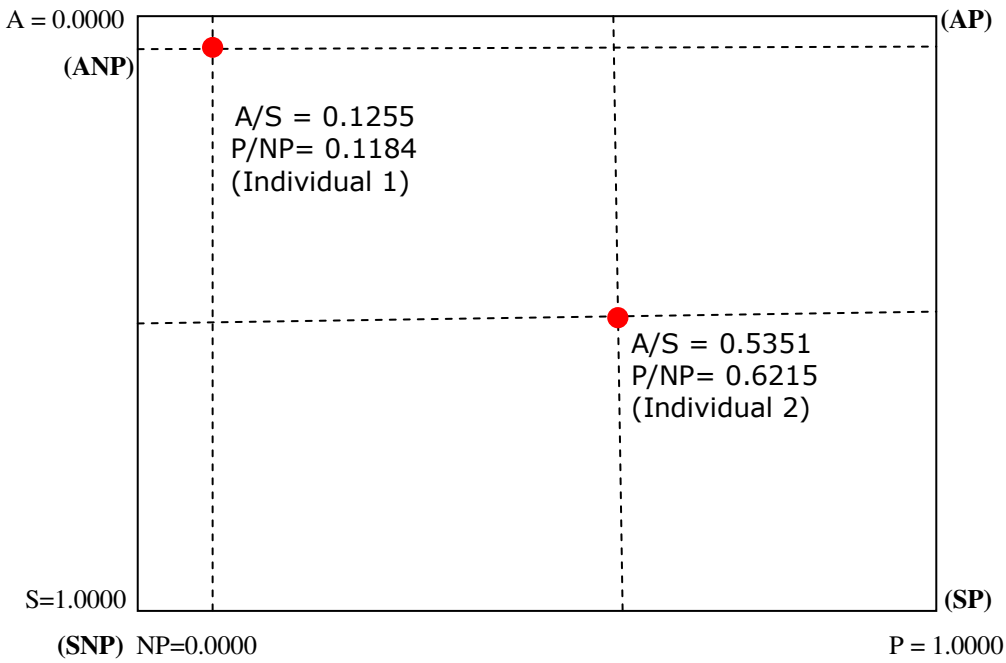
This section extends the two-person random interaction model (2PRIM) to incorporate strong reciprocity. However, before discussing this model, it is important to understand and clarify the two altruism-selfishness (A-S) and punishment-non-punishment (P-NP) continua that will be used. Bowles and Gintis (2000), Eldakar et al. (2007), Guth and Yaari (1992) and others have previously used discrete values of the altruistic/selfish behaviour (A-S) trait in their models. For example, Eldakar et al. (2007) has used the Nakamura and Iwasa (2006) framework that has four types of individuals: selfish punishers (SP), selfish non-punishers (SNP), altruistic punishers (AP) and altruistic non-punishers (ANP).

However in this thesis, 2PRIM uses two continua: altruism (A) – selfishness (S) and punishment (P) – non-punishment (NP), which have a range of 0.0000 – 1.0000 (extended to four decimals). Therefore, an individual with an A-S trait value of 0.0000 would be zero per cent selfish and 100 per cent altruistic. Similarly, if this individual has a P-NP trait value of 0.0000, this individual would have a zero per cent likelihood to punish (i.e. is a non-punisher). When both these continua are combined they can explain the four types of individuals in the Nakamura and Iwasa (2006) framework. For example, the selfish punisher (SP) will have the trait values – A-S: 1.0000 (100 per cent

selfish) and P-NP: 1.0000 (i.e. 100 per cent likely to punish). Similarly, the altruistic non-punisher (ANP) will have the trait values – A-S: 0.0000 (100 per cent altruistic) and P-NP: 0.0000 (i.e. zero per cent likely to punish).

In 2PRIM, punishment will be applied probabilistically. If an agent is less selfish than their opponent then they may punish the more selfish individual. The higher the individual's P-NP trait value the more likely that the individual will punish their opponent. In 2PRIM, individuals can have any combination of the A-S and P-NP continua. Two specific examples of two random individuals are provided in figure 6.1 that shows the first individual with an A-S trait value = 0.1255 and P-NP trait value = 0.1184. The second individual has an A-S trait value = 0.5351 and P-NP trait value = 0.6215. The corners of the box (in figure 6.1) represent the four types of individuals identified by Nakamura and Iwasa (2006).

Figure 6.1 Representation of 2PRIM A/S and P/NP continua



This two-person random interaction model (2PRIM) with strong reciprocity has been developed and simulated in *MATLAB R2008b* using largely the same process as in chapter 5. A pool of 1000 individuals is selected at random with the A-S and P-NP traits assigned to these individuals based on a uniform distribution. Each experiment comprises 1000 games and each game comprises 100,000 rounds. Two individuals are chosen at random for each round, and at the end of the round their utility is updated before they are sent back into the pool of 1000 individuals.

In the second phase of this round, the P-NP trait value shows the tendency of a less selfish individual punishing their opponent (who is more selfish). Punishment is applied when the P-NP value is greater than a probabilistic value (ascertained using a random number generator). The reason for this is so that punishment is not deterministic. Sometimes in every day interactions we will just let another's selfishness pass without reacting. At other times, and even in the same circumstances, we decide to stand our ground and respond. To account for such different behavioural response possibilities, the model requires an element of chance in this respect. In the event that punishment occurs, the punisher becomes more selfish by the Pfactor in the second phase of the round. However, this punisher (less selfish individual) will also need to pay a cost to punish because s/he is increasing his or her level of selfishness in the second phase to a level that is otherwise unnatural. Therefore, a punishment cost will be applied (i.e. Pcost), which will be deducted from the less selfish individual's utility if punishment is applied.

At the end of each game the 1000 individuals are sorted from highest to lowest utility, and the bottom 10 per cent of individuals are eliminated and replaced by new individuals. These new players have their A-S trait values assigned at random and the utility values of each player (in the pool of 1000 individuals) will be reset to zero at the

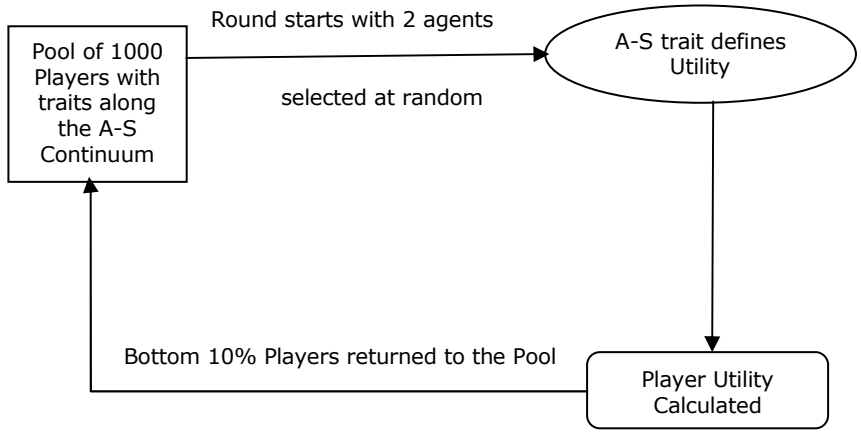
beginning of each game. This process will continue until the 1000 games are completed. Further detail about this game is provided in the Appendix. The process of eliminating the bottom 10 per cent of the individuals with the lowest utility at the end of each game continues to provide for the evolutionary process in this model. This evolutionary process helps to eliminate individuals who have lower levels of utility that were determined based on their A-S trait value in the first phase of each round and their A-S and P-NP trait values in the second phase of the round.

Utility (payoff) of each of these players will be calculated based on their A-S trait value using the equation below for the first phase of each round:

$$\frac{A_1 + A_2}{2} \cdot CGfactor + S_1 \cdot SFfactor - (S_1 + S_2) \cdot Cfactor \qquad (1)$$

Figure 6.2 explains Phase 1 (repeated for 100,000 rounds and 1000 games):

Figure 6.2 Diagrammatic representation of Phase 1 of the two-person random interaction model (2PRIM)



When comparing 2PRIM to models in Bowles and Gintis (2000), Fehr and Gächter (2002), Eldakar et al. (2007) and Eldakar and Wilson (2008), it is noticed that these models also have applied punishment with a cost that is borne by the punisher. This

punishment cost would look something like Pcost (i.e. described in 2PRIM), except that Pcost can be either a psychological or physical cost for being selfish. In Bowles and Gintis (2000), Eldakar et al. (2007) and others they represent it as a punishment cost that is incurred in eliminating other selfish individuals from the game. It is seen as a disutility from punishment. In Bowles and Gintis (2000), Fehr and Gächter (2002), Eldakar et al. (2007), Eldakar and Wilson (2008), there is no concept of Pfactor, where the level of selfishness increases for the less selfish player.

The role of Pfactor can have real-world parallels, for example, in real life you find altruistic people who will seemingly behave in a more selfish way in response to others' selfishness. For instance, if we are doing a work project together and the selfish individual contributes little effort to this work, I could complain to the manager in a manner that is 'out of character' for me. Alternatively, all the group members could vote out this selfish individual from this project and future projects (i.e. not want to work with this selfish individual). In contrast, a selfish individual may punish another selfish individual to improve their utility, as proposed by Eldakar et al. (2007) and Eldakar and Wilson (2008). For example, if I am a selfish individual in a group that has developed the next cutting edge technology, I might work with the group but try to get other selfish individuals eliminated before they can undermine the group's work and potential to benefit from the group's collective work.

Using the equation below each player's utility will be calculated at the end of the second phase of each round. The utility function now becomes:

$$\left(\frac{A_1 + A_2}{2} \cdot CGfactor\right) + \left(S_{X1} \cdot SFfactor\right) + \left(\left(S_{X1} + S_2\right) \cdot Cfactor\right) - \left(PCost\right) \qquad (2)$$

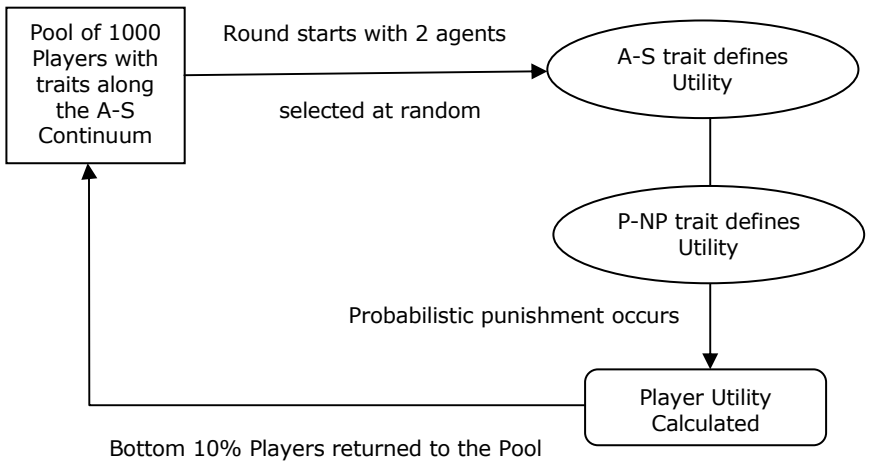
Where,

$S_{X1} = S_1 + Pfactor$ (where $0 < S_{X1} < 1$) (3)

Punishment is applied probabilistically in 2PRIM – where a random variable is used to assess if punishment will be applied. If this random variable (i.e. generated using a random number generator) is greater than the Pfactor value then punishment will be applied that will equal the value of Pfactor, else punishment will not be applied. This feature might be seen to add an element of realism. Often we might wish to punish but decline to do so. We might even try to do so but baulk at the last moment due to akrasia (i.e. shows weakness of will).

Figure 6.3 explains phase 1 and 2 of one round of each game after punishment has been incorporated in 2PRIM (this is repeated for 100,000 rounds and 1000 games):

Figure 6.3 Diagrammatic representation of rounds 1 and 2 of 2PRIM with strong reciprocity



At the end of this two-phase round, the utility of both these players will be updated based on equations 2-4 (provided above), and these two players will then be returned to the pool of the 1000 players with two new players then selected at random from this

pool to play the next round. Again, as these two players are selected at random some players could end up playing more rounds than others. This incorporates probabilities that occur with human beings in everyday life.

At the end of the game (100,000 rounds) the resource pool of 1000 players will be sorted by utility values in ascending order. The bottom 10 per cent of the players in this sorted list (by utility values) will be eliminated from the game. A same number (10 per cent) of new players will be accepted into the pool of 1000 individuals at random. Their A-S and P-NP values will be set at random, the utility values of each player (in the pool of 1000 individuals) will be reset to zero and the process will continue until the 1000 games are completed.

At the end of these 1000 games, results will show if the way I have modelled strong reciprocity in 2PRIM has had a noticeable impact on the average surviving levels of selfishness and utility. The next section will review the results of 2PRIM with strong reciprocity.

6.4 Results from this two-person random interaction model (2PRIM) with strong reciprocity

Results from the standard 2PRIM explain that selfishness increases and utility drops rapidly between games 1 to 100 (see Chapter 5). However, when punishment is applied to the standard 2PRIM, we find that the results change. I will use the 3-dimensional mesh graphs to perform this comparison. In particular I will compare the maxima in order to observe changes. Changes in maxima are not the only changes, but they give a good fast indication of the overall effects. We can also observe how the shape of the 3-dimensional curve alters.

In these simulations, the value of Pfactor was increased from 0.00 to 0.50, but no significant change in average the level of surviving selfishness or utility was observed. Further, simulations were conducted to keep the Pfactor constant at 0.50 and to increase Pcost from 0.00 to 1.00 (with Pcost values being 0.00, 0.10, 0.50 and 1.00). Results show that at Pcost = 0.00 (see figure 6.4 and 6.5), the level of selfishness and utility changed negligibly, compared to the standard 2PRIM, where no punishment is applied (see figures 5.8 and 5.12 in chapter 5).

The depression in the mesh in figure 6.4 is due to the increase in the level of selfishness that we see in figure 6.5 as the SFfactor increases from SFfactor = 0.4 to 0.6 (at CGfactor = 0.0). The level of selfishness did not significantly change, however. With the application of punishment a greater number of selfish individuals survived in subsequent games (see figure 6.5). These results are consistent with Bowles and Gintis (2000) and Fehr and Gächter (2002), where punishment results in an increase in altruism within the game. In comparison, these results are also consistent with results provided by Eldakar et al. (2007) and Eldakar and Wilson (2008), except that their models depict that the selfish punisher usually succeeds in punishing because selfish punishers can use utility provided by the altruists through the common pool (i.e. the altruistic punishers and non-punishers get less utility). In contrast, in 2PRIM the less selfish player obtains a higher utility, however still increasing the level of altruism within the game and decreasing the average level of selfishness.

Figure 6.4 Level of utility in 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.00, Cfactor = 0.25)

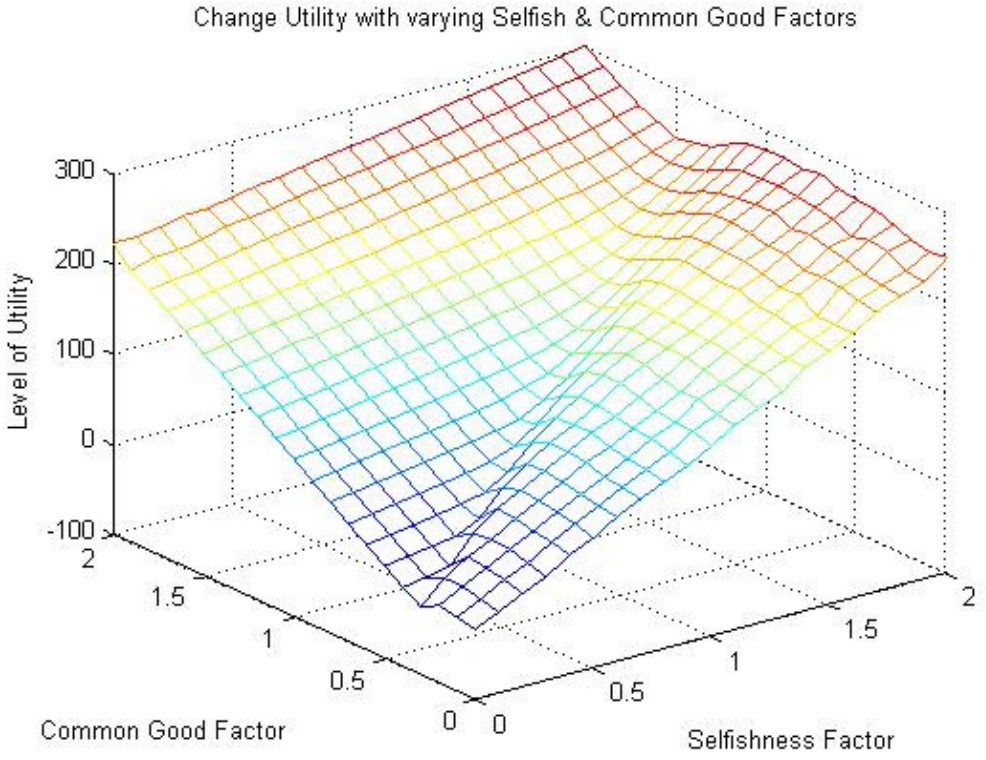
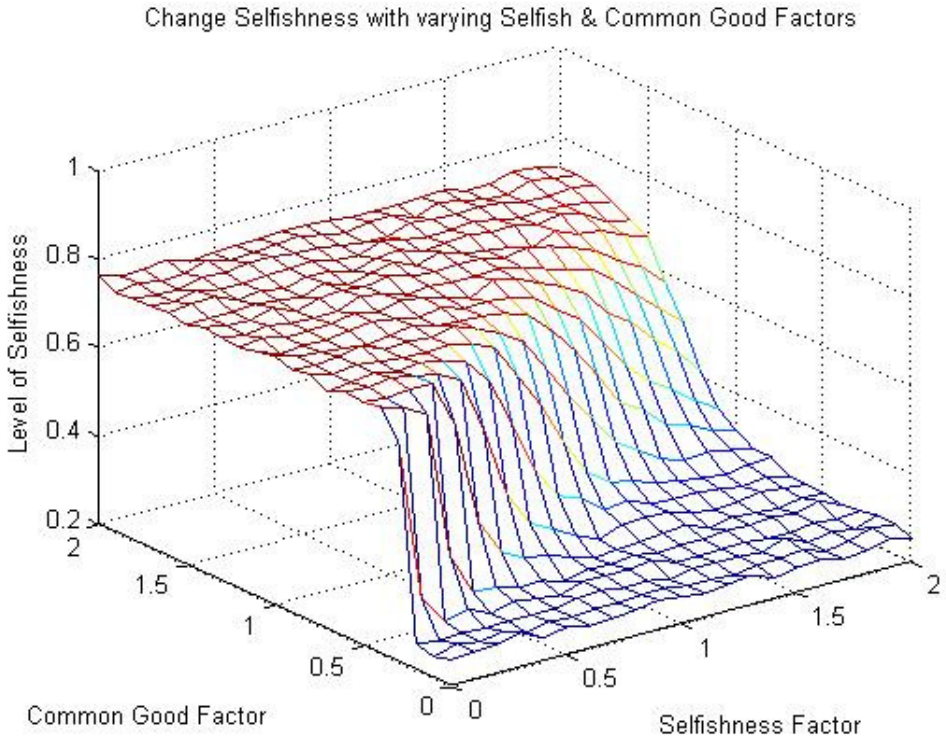


Figure 6.5 Level of selfishness in 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.00, Cfactor = 0.25)



However, if an altruistic punisher tries to punish other individuals, they will have even lower utility, as they are unable to recoup utility like the selfish individuals. As a result, they tend to be removed from the game. Bowles and Gintis (2000) call strong reciprocity second-order altruism because they believe that punishment increases the surviving level of altruism, while Eldakar and Wilson (2008) call it second-order altruism because the selfish individuals punish to eliminate other selfish individuals to improve their future utility. Regardless, this encourages an increase in altruism due to the selfish punisher implementing punishment. 2PRIM by comparison, provides for more altruistic to become more selfish in the second phase of each round. In both cases, the number of altruists surviving increases with an increase in the level of punishment.

When the Pfactor is kept constant at 0.50, but the Pcost is increased to 0.10, no significant change is noticed (see figures 6.6 and 6.7) compared to Pfactor = 0.50 and Pcost = 0.00 (see figures 6.4 and 6.5). This shows that a marginally increase in Pcost (cost for punishment borne by the punisher) has minimal if any impact on the level of selfishness or utility in 2PRIM with strong reciprocity.

The results initially seem consistent with those provided by Bowles and Gintis (2000), Boyd et al. (2003) and Eldakar and Wilson (2008) due to the fact that there has only been a marginal increase in Pcost and it has not had a significant impact of the utility of the less selfish individual. However, the difference between 2PRIM and these models is that a significant enough increase in Pcost results in a greater number of the altruists being eliminated. These other models do not have a concept of Pcost. Therefore this loss of utility due to the increase in Pcost is not seen in other models. This difference can be seen as Pcost is increased in figures 6.6 – 6.11.

Figure 6.6 Level of utility in 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.10, Cfactor = 0.25)

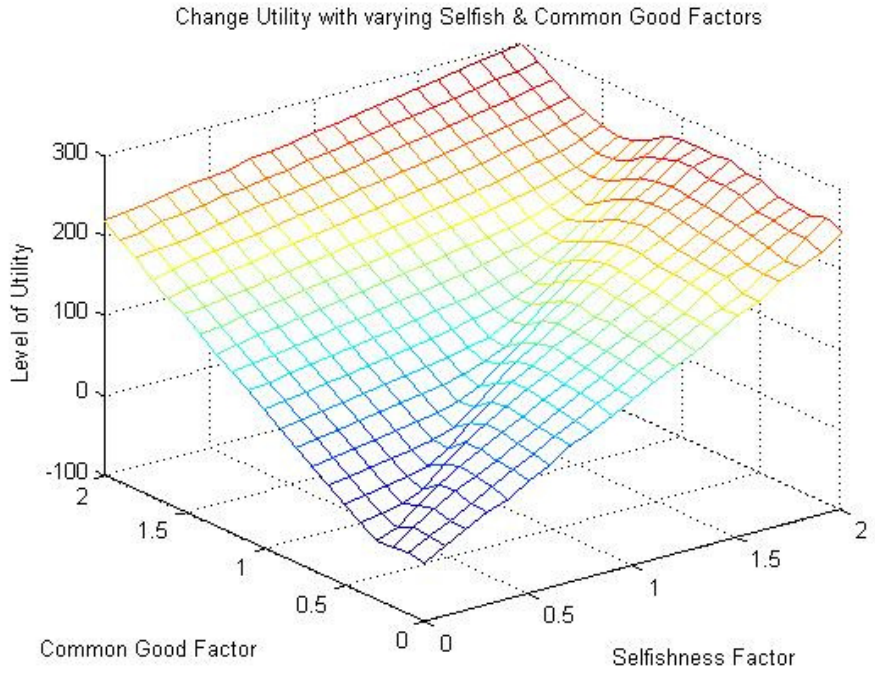


Figure 6.7 Level of selfishness in 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.10, Cfactor = 0.25)

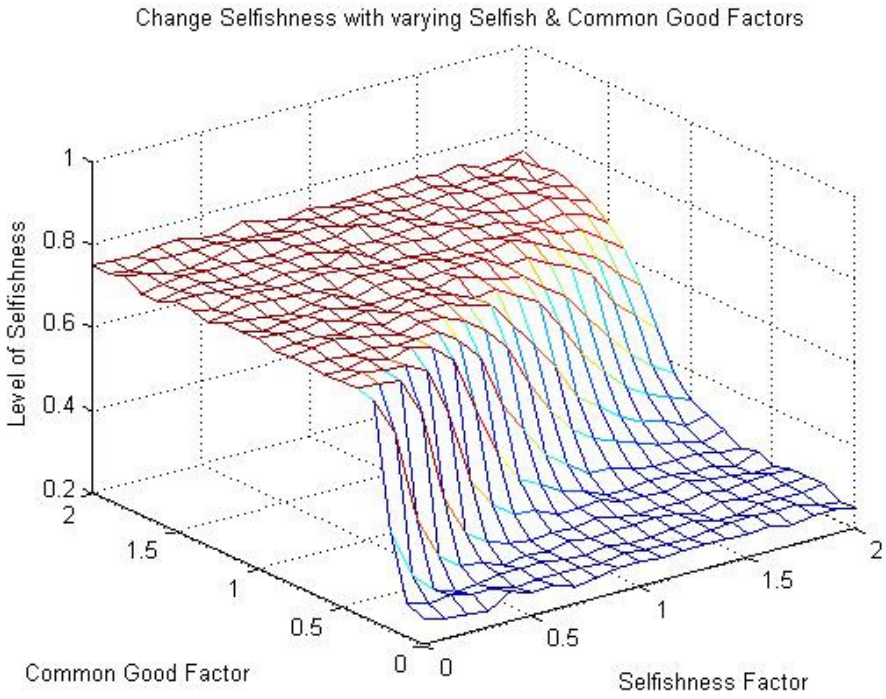


Figure 6.8 Level of utility in 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.50, Cfactor = 0.25)

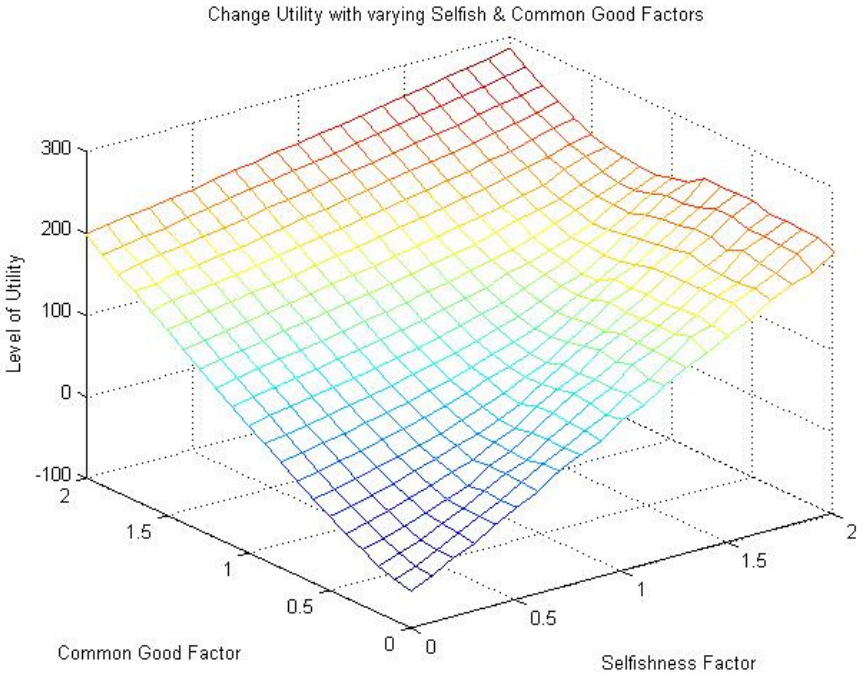


Figure 6.9 Level of selfishness 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.50, Cfactor = 0.25)

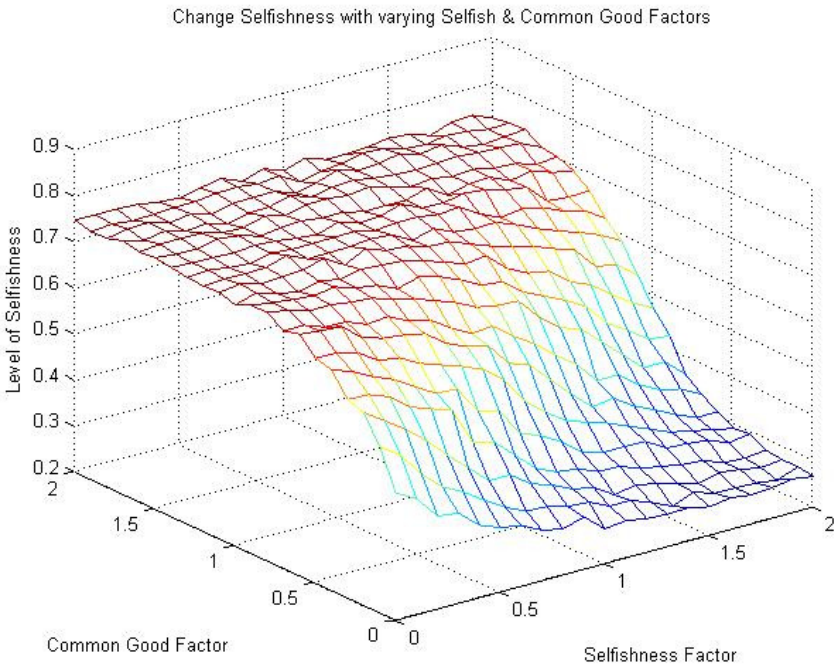


Figure 6.10 Level of utility in 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.75, Cfactor = 0.25)

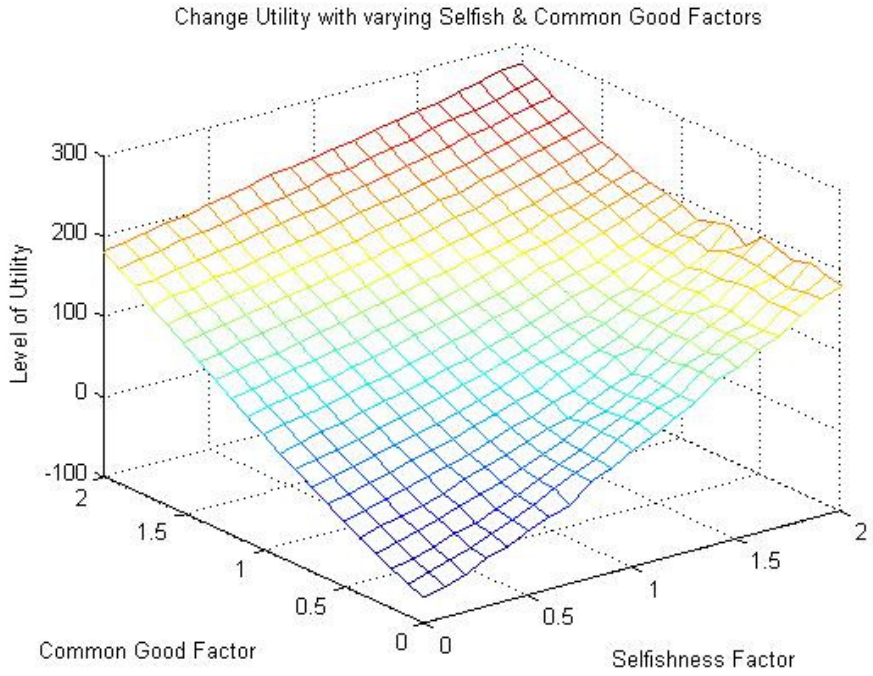
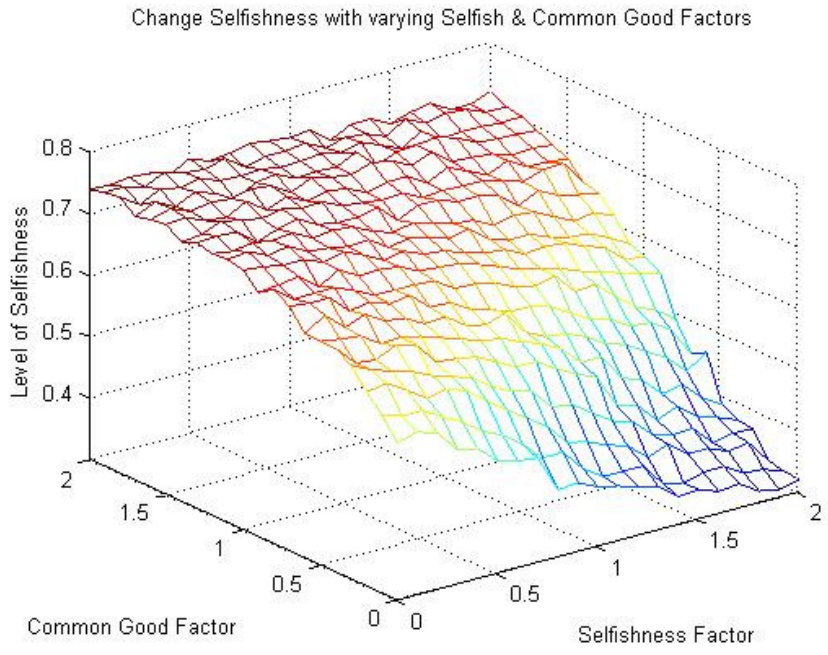


Figure 6.11 Level of selfishness 2PRIM with strong reciprocity (Pfactor = 0.50 and Pcost = 0.75, Cfactor = 0.25)



When, Pcost is increased to 0.50 (as in figure 6.9), it is noticed that the selfishness mesh becomes smoother. This result is due to the fact that the cost to the less selfish individuals has increased due to punishing. This in turn has reduced the number of less selfish individuals. It has also had an impact on the more selfish individuals, resulting in the removal of the phase transition (that occurred in figure 6.7). As some altruistic (less selfish) individuals are still surviving, they help smooth out the selfishness mesh in figure 6.9. In Eldakar and Wilson (2008), the utility of each selfish punisher reduces, but they still dominate the population as they have obtained utility in the initial round of the game by acting selfishly. However, the results in this thesis show that majority of the less selfish individuals are actually being eliminated from subsequent games. Hence, the overall utility mesh in figure 6.8 is unchanged.

As Pcost is increased to 1.00, the selfishness mesh (see figure 6.11) becomes flatter, due to the fact that fewer altruistic individuals exist in this game. This can be seen as the utility mesh (see figure 6.10) changes marginally compared to Pcost = 0.50 (see figure 6.8). This shows that Pcost is having a significant impact on the possibility of altruistic individuals to survive. Eldakar and Wilson (2008), Bowles and Gintis (2000) and Boyd et al. (2003), explain that an increase in the level of punishment should increase the level of altruism in the game. However, the results of 2PRIM with strong reciprocity offer different possibilities. This occurs as the negative impact of the punishment cost (Pcost) is being borne by the more altruistic (less selfish) individuals rather than the more selfish individuals as in the case of these other models. That is, it reduces the utility of the altruistic individuals in future games and results in them being eliminated.

A real life example would be where a selfish person overtakes an altruistic individual on the road and the altruistic individual shouts at the selfish individual. However, the altruistic individual then feels remorse for shouting at the selfish individual (i.e. resultantly facing a psychological cost for shouting). This psychological cost is an instance of negative utility that

applies to the more altruistic individual. As this altruistic individual feels remorseful, it could in reality affect other activities this person may perform after this incident. On the other hand, if this incident did not occur then this psychological cost would not have been borne by the altruistic individual.

We have analysed the way strong reciprocity affects the change in the level of selfishness and utility. It is now also important to compare the impact of strong reciprocity with the impact of the return on common good (CGfactor), selfish investment (SFfactor) and the cost of competition (Cfactor). Previously, Bowles and Gintis (2000), Eldakar and Wilson (2007), Eldakar (2008), Boyd et al. (2003), Fehr and Gächter (2000) and others have shown that strong reciprocity has a positive impact on altruism within groups. The results in this thesis show that strong reciprocity does not significantly increase the level of selfishness or utility in 2PRIM. However, an increase in the cost of punishment (Pcost) does have an impact of reducing the number of altruistic individuals in subsequent games.

However, if the levels of the CGfactor, SFfactor or Cfactor are altered, the overall dynamics of the game change by constricting the pay-offs to each player on the common (public) good, selfish investment and through the cost of competition. In 2PRIM, while punishment has a marginal effect, it is noticed that changes in the common good, selfish and cost of competition factors have a significant effect.

An increase in the common good factor (CGfactor) to levels greater than 1.00 reduces the level of selfishness within the game and increases the overall level of utility more than does a change in the level of strong reciprocity (Pfactor or Pcost). This happens as strong reciprocity only assists some altruists survive as they obtain higher utility due to the probabilistic punishment which only increases the level of selfishness for a few altruists. However, an increase in the CGfactor provides an increase in the utility of all altruistic individuals that is

higher than the increase in utility for the selfish individuals, resultantly assisting a greater number of altruists to survive.

In contrast, an increase in the selfish factor (SFfactor) to levels greater than 1.00 will increase the level of selfishness and utility within the game. Here, too the effect is more than is a change in the level of strong reciprocity. A higher SFfactor assists more selfish individuals survive, but a decrease in the level of guilt (lower SFfactor) results in more altruists surviving (as utility decreases for all selfish individuals) compared to the application of strong reciprocity, which assists only a few altruists as only a few altruists will have sufficient utility to survive against selfish individuals within the game.

When, the Cost of competition (Cfactor) has a value greater than 0.25, simulations show that the level of selfishness within the group reduces within subsequent games. This happens due to the fact that a higher Cfactor reduces the utility of selfish individuals who compete with other selfish individuals in the two-person game, causing a reduction in utility due to destructive competition. Strong reciprocity (an increase in Pfactor) temporarily increases the selfishness of altruists in order to increase their utility (or reduces utility when Pcost is higher than Pfactor). However, an increase in the Cfactor can result in the more selfish individuals being eliminated.

Finally, the analysis provided through the simulations of 2PRIM with strong reciprocity shows that it would be better to change the levels of the common good (CGfactor), selfish (SFfactor), level of guilt or the cost of competition (Cfactor) factors rather than apply strong reciprocity to 2PRIM. Strong reciprocity has a smaller impact on changing the level of selfishness or utility compared to these factors in 2PRIM.

6.5 Conclusion

This chapter develops 2PRIM further by adding the concept of strong reciprocity over and above the concepts of the common good, selfish and cost of competition factors that we discussed in the previous chapter. Strong reciprocity is applied in this model by temporarily increasing the level of selfishness of the less selfish (altruistic) individual in second phase of each round. Punishment is applied on a probabilistic basis, where the punishment factor is greater than a random number. Results show that when punishment is applied in this model, the level of selfishness and utility marginally decreases. These results are consistent with Bowles and Gintis (2000), Boyd et al. (2003), Eldakar and Wilson (2007) and Fehr and Gächter (2002), in that an increase in punishment results in a higher level of altruism. In 2PRIM, as altruistic individuals become temporarily more selfish, it assists a greater number of altruistic individuals to survive. When the cost of punishment is increased in 2PRIM the number of altruistic individuals decreases, as the cost of punishment outweighs the increase in utility from being temporarily more selfish. In these other models, a concept similar to Pcost does not exist.

In comparison to results in Chapter 5, it is seen that an increase in the common (public) good, selfish or the cost of competition factors could likely have a longer-lasting effect on the levels of selfishness and utility compared to strong reciprocity (punishment). This occurs as these three factors compared to punishment provide incentives to change pay-offs for individuals playing the game, while punishment only provides temporary assistance to altruists in the second phase of each round if probabilistic punishment is implemented. Additionally, as discussed, if the Pcost increases, punishment has even less effect in increasing altruism within a game.

CHAPTER SEVEN

Conclusion

7.1 Summary

This thesis poses the question: if we model selfishness, altruism, reciprocity, competition and the common or public good in such-and-such ways, what might be the individual, social and evolutionary consequences of everyday two-person random interactions? In order to undertake these tasks, the thesis starts by discussing Adam Smith's approach to self-interest and altruism, which he formulated as parts of a comprehensive theory in his two books, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776) and *The Theory of Moral Sentiments* (1790). Chapter 2 focuses on Smith and the insights that we can still fruitfully extract from his seminal works. It also discusses the concepts of selfishness and altruism in the contexts of associated research areas: philosophy, socio-biology, behavioural economics and neuroeconomics. This chapter also defines the concepts of selfishness and altruism for use throughout this thesis. The discussion of Smith also is a stand-alone contribution of this thesis, since it helps to remedy one-sided views that still prevail. For instance, Smith was not an advocate of selfishness, but a subtle thinker who emphasised an appropriate mix of the virtues of prudence, benevolence, justice, duty and self-command.

In chapter 3, concepts of selfishness and altruism are reviewed further in terms of the work of Bowles and Gintis (2004). They have coined the term 'strong reciprocity'. Strong reciprocity can be defined in their terms as second-order altruism, whereby an individual, at a cost to themselves, punishes selfish individuals who do not follow the group's social norms (Bowles and Gintis 2004).

In chapter 4, game theory and complex systems related research was been reviewed extensively, as it provides the methodology of the model that the thesis develops in chapters 5 and 6. Chapter 4 briefly discusses the interlinkages between game theory, behavioural economics and social cost. It also analyses how complex systems relate to the concept of strong reciprocity, specifically looking at the models by Bowles and Gintis (2000), Boyd et al. (2003) and Eldakar et al. (2007) in depth to provide a basis for the discussion in chapter 5. The model developed in that chapter takes its inspiration from the models in these papers.

In chapter 5, the two-person random interaction model (2PRIM) is proposed. It is a two-person Prisoners' Dilemma model developed using behavioural trait values of altruism and selfishness. This model is developed as described in Chapter 1.3 in order to simulate individual, social and evolutionary consequences of everyday two-person random interactions. In effect, 2PRIM uses variables for selfishness, altruism, reciprocity, competition and the common or public good in certain ways in order to model possible individual, social and evolutionary consequences of everyday two-person random interactions.

Recall that Chapter 1.5 stated that the thesis has modest objectives. Based on the a comprehensive discussion of the above concepts and the literature surrounding them, the purpose of the thesis is to offer a modelling approach that those who are endeavouring to answer the big questions can use as a tool. The modelling approach the thesis contributes asks, if human selfishness, altruism, reciprocity, competition and the common or public good were found to be such-and-such and related to each other with such-and-such weights attached to them, what might be the individual, social and evolutionary consequences of everyday two-person random interactions? It then tries to analyse if the evolutionary level of altruism increases with a change in the return on the common pool (public good) investment, selfish investment or the cost of competition.

In chapter 6, the 2PRIM model is extended to include punishment (strong reciprocity) and the idea is to understand if the application of punishment will increase the surviving level of altruism within the game. Bowles and Gintis (2000), Boyd et al. (2003), Eldakar et al. (2007), Fehr and Gächter (2007) and Gintis et al. (2008) have developed models that incorporate strong reciprocity. In general, these models show that punishment has an impact of increasing the level of altruism within groups. This occurs primarily due to the threat of punishment or ostracism that results in selfish individuals working in co-operation with the group to avoid being removed in subsequent games. Researchers have not been able to explain other reasons for the increase in altruism within groups even in one-off games. Some suggestions that have been made in relation to an increase in altruism within groups point to research in the areas of genetic co-evolution, kin selection as well as social and cognitive factors. Strong reciprocity has had strong support in explaining an increase in the level of altruism within groups based on genetic co-evolution.

This thesis extends this paradigm by understanding if it is possible to increase altruism by changing the pay-offs for common (public) good investments, selfish investments and the cost of competition rather than just to understand the impact of punishment in this game. This thesis finds that an increase in the return on common good investment has a positive impact on the level of altruism and increases the mean utility within the game. An increase in the return on the selfish investment has a positive impact on increasing the average level of surviving selfishness and this also increases the mean utility within the game. However, the mean utility in this case does not increase as much as it does when the return on the common good investment is increased. An increase in the cost of competition also has a significant impact on the level of selfishness within the game, as it reduces the utility of selfish individuals when they interact with other selfish individuals. The mean utility in the game also reduces as the selfish individuals obtain negative utility due to the cost of competition.

Punishment is applied on a probabilistic basis in the expanded 2PRIM model. But, when punishment is applied, the average level of surviving selfishness marginally decreases. Any further increase in the level of punishment has minimal impact on the level of selfishness and utility within this game. Punishment in this model is applied by the less selfish (altruistic) individual becoming temporarily more selfish within the second phase of each round. This is akin to Bowles and Gintis (2000) but differs from Eldakar et al. (2007), where the more selfish individuals punish other equally or less selfish individuals. However, there is also a cost to punish other individuals in the 2PRIM model, and this cost is borne by the less selfish individual (i.e. the one who punishes).

This cost of punishment is a psychological or physical cost borne by the less selfish individual becoming temporarily more selfish in the 2PRIM. Resultantly, an increase in this cost of punishment has a negative impact on the utility of the less selfish individual and the average level of surviving altruism decreases within the game. Effectively, when more altruistic individuals face a higher cost of punishment, they are more likely to obtain lower utility and may subsequently be removed from the game. This thesis also shows that punishment can only temporarily change the level of selfishness and utility within the game. However, when the return on common (public) good investment, selfish investment and cost of competition are changed, then the pay-offs change on a more permanent basis. Also, the change in these three factors has a more significant effect in 2PRIM than does punishment on the level of altruism and utility within the game.

7.2 Contribution to knowledge

In this thesis, I have reviewed research across the different areas of economics, behavioural economics, neuroeconomics, philosophy and socio-biology to analyse the concepts of

selfishness, altruism and strong reciprocity in chapters 2-4. The literature review supported the development of a random, one-shot or non-repeated (random) interaction model as a significant contribution to knowledge. A computer simulation model called the two-person random interaction model (2PRIM) was developed to assist in understanding such interactions. This 2PRIM model was used to evaluate the effects of selfishness, altruism and strong reciprocity on the consequences of two-person random interactions within a Prisoners' Dilemma structure. In 2PRIM, selfishness and altruism were defined on a continuum instead of being represented as discrete values, as had been the case in previous research, for example, Bowles and Gintis (2000) and Eldakar et al. (2007). The use of defining selfishness and altruism on a continuum was that a much larger variety of individual traits could be noticed, and that reflects the existence of similar traits in the everyday human population in a more realistic fashion.

While other scholars have developed two-person models in the past, none of them have analysed two-person random interactions where selfishness and altruism has been defined on a continuum. The closest parallels are those of Eldakar et al. (2007) and Eldakar and Wilson (2008). Here Eldakar et al. (2007, p. 199) stated that trait values in their N-person interaction model were allocated in a way 'that initially varied uniformly ... between 0 and 1 at 0.1 increments'. However, the number of players 'N' is always greater than two. Further, the model in Eldakar and Wilson (2008, p. 6982) explained that it allocated traits or behaviours for selfishness, altruism and strong reciprocity as 'pure strategies', which meant that these players were either pure altruists or purely selfish and either strong reciprocators or not. Earlier work on two-person games also has not included random interaction experiments (2008, p. 6982, citing Axelrod 1984, Hamilton 1964, 1975, Axelrod and Hamilton 1981, Maynard Smith 1982) and have not considered strong reciprocity.

The first contribution of this thesis was to model a *significant* problem that scholars seek to understand, being the random, one-shot or non-repeated human interactions. We

face such interactions in nearly everything we do in our daily lives, and Silk (2005, pp. 63-4) pointed out that such interactions are of evolutionary significance.

The second contribution of this thesis was to model a specific set of random interactions, being those of two persons, continuous-trait attributes and strong reciprocity. The third contribution was that this thesis created an original game-theoretic computational model (2PRIM) in order to contribute to the understanding of the individual, social and evolutionary consequences of everyday two-person random interactions in defined circumstances. Those circumstances were structured according to the standard Prisoners' Dilemma framework, which constrains the practicality of results generated by 2PRIM to circumstances (i.e. random two-person interactions) that correspond to Prisoners' Dilemmas. Given that many circumstances do correspond, and given the common role of Prisoners' Dilemmas in economics and game theory, 2PRIM can contribute originally to our knowledge of human interaction and its consequences (see the fifth and final contribution below).

The fourth contribution was that this thesis contributed originally to knowledge by developing the foregoing contributions with the context of a theoretical discussion of the literature – including the work of Adam Smith – of the emerging multidisciplinary field of research. In is important to note, that only within this context could the questions posed in this thesis would possibly be understood.

The final contribution of this thesis was that it explained in chapter 6 that pay-offs for players could change by varying the selfish factor (SFfactor), common good factor (CGfactor) and cost of competition factor (Cfactor). This could possibly constitute a substitute to applying strong reciprocity within random interactions in Prisoners' Dilemma contexts (see contribution three above).

7.3 Limitations to this research

This thesis has tried to understand two-person random interactions in an everyday setting from a theoretical perspective and it has certain limitations that are explained below:

1. *Psychological and game theoretic limitations:* This thesis only models selfishness, altruism and strong reciprocity in a two-person Prisoners' Dilemma context. It does not intend to include any other psychological or game theoretic framework/concept. Nor does it pursue comprehensive theories of morality, human nature and the evolution of altruism and selfishness. Indeed, to model such comprehensive theories might mean setting aside the Prisoners' Dilemma framework altogether.

2. *Complex systems limitations:* Complex systems attempt to explain the complex inter-relationships that exist in society. When attempting to analyse two-person random interactions, the complexity involved in representing and analysing such interactions can be difficult due to the existence of other factors affecting such interactions. This thesis has tried to model two-person random interactions in a more simplistic way, primarily by only concentrating on these two-person interactions between strangers. In society, however, as human beings we interact with other people closer to us (for example, friends, acquaintances and family), which can alter our interactions with strangers over time due to the learning we obtain from these other interactions and relationships. This learning would change the dynamics of or could alter the results of the two-person random interaction experiment over time. Individuals start to learn and behave differently. Additionally, other complex systems limitations could apply as the level of complexity (affecting both the computational and game theoretic model) incorporated in 2PRIM has been limited.

3. *Absence of learning:* No learning is expected to take place between individuals throughout the games modelled in 2PRIM. Individuals only punish based on their probabilistic punishment trait in this model and this is not associated to learning as commonly understood in game theory.

4. *Limitations related to the review of Adam Smith's work:* There has been extensive research undertaken in relation to Adam Smith's work. This thesis has covered the broad range of this research. However it has concentrated more on Smith's work itself rather than secondary or integrative sources.

5. *Comparative realism of assumptions:* In order for the 2PRIM model to be used in practical (or real-world) situations, its structure (e.g. the Prisoners' Dilemma constraint) and outcomes must align with what might have been or might be important evolutionary scenarios (see e.g. Eldakar and Wilson's examples referred to on p. 139 above). A set of comparisons of what 2PRIM allows versus other scholars' models must be undertaken before 2PRIM can stand to make predictions or assessments of evolutionary situations. Some possibilities are suggested in section 7.4.

6. *Limitation of the application of selfless altruism (i.e. Buddhist Economics and Gandhism):* It is possible that some altruists sacrifice their wellbeing for the greater good of others. However, the idea of this thesis is to specifically review the research problem that deals with two-person random interactions. Therefore, the concept of selfless altruism, which would include Buddhist Economics and Gandhism, will be explicitly excluded from this thesis, except in the rare case when A =1 and S = 0 as randomly selected.

7.4 Possible applications of this research

2PRIM can be applied to the following situations:

1. In a small-world network to understand human behaviour in random interactions in society to understand how to better utilise social networks. For example, Facebook and MySpace (social-networking technology) can be improved when random social interactions are modelled using 2PRIM.

2. To explain evolution of social behaviour in non-repetitive random interaction games to help improve marketing/advertising to include social evolution in such interactions. 2PRIM in a small-world framework can help to possibly simulate random, semi-permanent and permanent interactions between humans that can be used to help us to adapt marketing and financial products from a social-interaction viewpoint.

3. Understanding the social evolution and regulation patterns of genes (within a small world) to understand how genes might evolve in a network through random interactions.

4. In a small-world network with limited information, to explain financial market price dynamics as investors in markets undertake random interactions in buying/selling assets with incomplete information through floor-traded and over-the-counter exchanges.

5. To help to simulate interactions and possible outcomes in negotiations between two parties who have not interacted with each other previously and have met for a one-off negotiation.

6. Improving altruism and reducing social cost: This model shows that changing the rate of return on the common good investment (CGfactor), the Guilt factor or the Cost of Competition (Cfactor) will have a greater impact on increasing altruism in society compared to an increase in strong reciprocity. This could have applications in a number of societal interactions, for example, reducing environmental impact by lowering pollution where models can incorporate different human behavioural choices through changes in pay-offs for common good and selfish investments rather than applying strong reciprocity.

7.5 Suggested extensions

After, discussing the limitations to this thesis, it is important to also consider what future research could be developed from this model.

While two-person human interactions in everyday activities are commonplace, these activities are undertaken by us as part of a society. Therefore, the possibility of developing this research into the concept of two-person human interaction in networks or small worlds is promising. We as humans have relationships (permanent or semi-permanent) with other individuals in society, for example our immediate or extended family (permanent relationships) and friends or acquaintances (semi-permanent or temporary relationships). Also, we do not know everyone in the world. Therefore these networks will consist of individuals we do not know and those we do know. Effectively this would lead to the construction of a 2PRIM in a small-world network to begin with. Individuals would possess these four types of individual relationships: permanent, semi-permanent or temporary, random and no relationship at all.

Another extension of 2PRIM might be to help include genetic learning, where only successful individuals survive and reproduce, while the less successful individuals are removed from subsequent games. In this way, the replicator dynamic in this model will support the

heritability of successful strategies. In addition, while genetic learning is important in this model, it is also important to explain individual learning in society. As 2PRIM currently deals with unrelated individuals in random one-off games, it might be adapted so that individuals learn through punishment in games with strangers. For example, 2PRIM might model each individual's learning from these interactions at different rates. Thus 2PRIM might be adapted to depict the way humans learn and change their behavioural responses (traits).

There is a possibility to extend the 2PRIM in a way that punishment can be remodelled to be applied in other ways in comparison to the way it is applied in the existing 2PRIM. This could provide further variations to the way punishment is applied and to make the application of punishment more realistic (i.e. as it would happen in everyday two-person random interactions).

We could also analyse the different impacts of factors related to fairness, trust, emotions, beliefs, intentions and fear in such two-person human interactions. Such analysis could provide a basis for simulating interaction in society as a whole. It could provide a view of interactions that we undertake with people who are unrelated to us, and it seems plausible that such interactions may form most of the transactions we undertake in our everyday life. This would provide us a better understanding of how different factors within the society change in such relationships. The changes required to 2PRIM might include factor changes (e.g. addition additional factors) or modifications to existing factors (as SFfactor was modified by adding punishment). In doing so, we might reconsider the usefulness of the Prisoners' Dilemma framework (2PRIM parameters), as suggested at section 7.3(1).

In conclusion, while these are a few possible avenues of future research using 2PRIM, there could be other variations that could be used to develop this model further. This, in turn, could assist economics to use evolutionary game-theoretic approaches to understand better the

nature and evolution of altruism and selfishness – two concepts at the heart of human social interaction.

REFERENCES

Akerlof, G. A. (1982). Labor Contracts as Partial Gift Exchange. *The Quarterly Journal of Economics* **97**(4): 543-569.

Akerlof, G. A. and J. L. Yellen (1988). Fairness and Unemployment. *The American Economic Review* **78**(2): 44-49.

Akerlof, G. A. and J. L. Yellen (1990). The Fair Wage-Effort Hypothesis and Unemployment. *The Quarterly Journal of Economics* **105**(2): 255-283.

Alexander, J. M. (2005). The Evolutionary Foundations of Strong Reciprocity. *Analyse & Kritik* **27**(1): 106-112.

Alison, G., L. R. McQueen and L. M., Schaerfl (1992). Social Decision making processes and the Equal Partitionment of Shared Resources. *Journal of Experimental Social Psychology* **28**:23-42.

Anderson, P. W., K. Arrow and D. Pines. Eds. (1988). *The Economy As An Evolving Complex System.* Westview Press, Colorado.

Andreoni, J. (1988). Privately provided public goods in a large economy: The limits of altruism. *Journal of Public Economics* **35**(1): 57-73.

Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving? *Economic Journal* **100**(401): 464-477.

Andreoni, J. (1993). An Experimental Test of the Public-Goods Crowding-Out Hypothesis. *American Economic Review* **83**(5): 1317-1327.

Andreoni, J. (1995). Cooperation in Public-Goods Experiments: Kindness or Confusion? *American Economic Review* **85**(4): 891-904.

Arthur, W. B., S. N. Durlauf and D. Lane. Eds. (1997). *The Economy As an Evolving Complex System II.* Perseus Publishing, New York.

Arrow K. J. and G. Debreu (1954). The Existence of an Equilibrium for a Competitive Economy. *Econometrica* **22**:265-290.

Ashraf, N., C. F. Camerer and G. Loewenstein (2005). Adam Smith, Behavioural Economist. *The Journal of Economic Perspectives* 19(3): 131-145.

Aumann, R. J. and J. Drèze (1974) Cooperative Games with Coalition Structures. *International Journal of Game Theory* **3**: 217–237.

Axelrod, R. M. (1984). *The Evolution of Cooperation.* Basic Books, New York.

Axelrod. R. M. and W. D. Hamilton (1981). The Evolution of Cooperation. *Science* **211**(4489): 1390-1396.

Batson, C. D. (1991). *The Altruism Question: Toward a Social Psychological Answer*. Routledge, New York.

Batson, C. D. and L. L. Shaw (1991). Evidence for Altruism: Toward a Pluralism of Prosocial Motives. *Psychological Inquiry* **2**(2): 107-122.

Bazerman, M. H. and M. A. Neale (1983). Heuristics in negotiation: Limitations to effective dispute resolution. In *Negotiating in Organizations*. M. H. Bazerman and R. J. Lewicki. (eds). Sage Publications, Beverly Hills.

Beaulier, S. and B. Caplan (2007). Behavioural Economics and Perverse Effects of the Welfare State. *Kyklos* **60**(4): 485-507.

Bechlivanidis, C. (2006). An examination of the role of prestige in cultural evolution with the aid of Agent Based Modelling. *Computer Science*. Bath, United Kingdom, University of Bath. **BSc (Hons) Computer Science:** 98.

Becker, G. S. (1976). Altruism, Egoism and Genetic Fitness: Economics and Sociobiology. *Journal of Economic Literature* **14**(3): 817-826.

Becker, G. S. (1981). Altruism in the Family and Selfishness in the Market Place. *Economica* **48**(189): 1-15.

Becker, G. S. (1993). Nobel Lecture: The Economic Way of Looking at Behaviour. *The Journal of Political Economy* **101**(3): 385-409.

Benoit, H.-V. (2007). Decision-Making: A Neuroeconomic Perspective. *Philosophy Compass* **2**(6): 939-953.

Berg, J., J. Dickhaut and K. McCabe (1995). Trust, Reciprocity, and Social History. *Games and Economic Behaviour* **10**(1): 122-142.

Bergstrom, T. C. and O. Stark (1993). How Altruism Can Prevail in an Evolutionary Environment. *The American Economic Review* **83**(2): 149-155.

Bernheim, B. D., A. Shleifer and L. H. Summers (1985). The Strategic Bequest Motive. *The Journal of Political Economy* **93**(6): 1045-1076.

Bester, H. and W. Guth (1998). Is altruism evolutionarily stable? *Journal of Economic Behaviour & Organization* **34**: 193-209.

Bewley, T. F. (1999). Why wages don't fall during a recession. Harvard University Press, Cambridge.

Binmore, K. G. (1998). *Just Playing: Game Theory and the Social Contract*. MIT Press, Cambridge.

Bittermann, H. J. (1940). Adam Smith's Empiricism and the Law of Nature. *The Journal of Political Economy* **48**(5): 703-734.

Blume, L. E. and S. N. Durlauf. Eds. (2005). *The Economy As an Evolving Complex System III: Current Perspectives and Future Directions*. Oxford University Press, New York.

Boehm, C., H. B. Barclay, R.K. Dentan, M. C. Dupre, J. D. Hill, S. Kent, B. M. Knauft, K. F. Otterbein and S. Rayner (1993). Egalitarian Behaviour and Reverse Dominance Hierarchy [and Comments and Reply]. *Current Anthropology* **34**(3): 227-254.

Bolton, G. E., J. Brandts and A. Ockenfels (1998). Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game. *Experimental Economics* **1**(3): 207-219.

Bolton, G. E. and A. Ockenfels (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review* **90**(1): 166-193.

Boulding, K. E. (1969). Economics as a Moral Science. *The American Economic Review* **59**(1): 1-12.

Bowles, S. (2000). Economic institutions as ecological niches. *Behavioural and Brain Sciences* **23**(01): 148-149.

Bowles, S. (2003). *Microeconomics: Behavior, institutions, and evolution.* Princeton University Press, Princeton.

Bowles, S. (2008). Being Human: Conflict: Altruism's midwife. *Nature* **456**:326-327

Bowles, S. (2008). Policies Designed for Self-Interested Citizens May Undermine The Moral Sentiments: Evidence from Economic Experiments. *Science* **320**(5883): 1605-1609.

Bowles, S. and H. Gintis (2000). The Evolution of Reciprocal Preferences. Santa Fe Institute, New Mexico**:** 1-20.

Bowles, S. and H. Gintis (2002). Social Capital and Community Governance. *The Economic Journal* **112**(483): F419-F436.

Bowles, S. and H. Gintis (2002a). Behavioural science: Homo reciprocans. *Nature* **415**(6868): 125(4).

Bowles, S. and H. Gintis (2002b). The Origins of Human Cooperation. In *Genetic and Cultural Evolution of Cooperation*. P. Hammerstein. Ed. MIT Press, Cambridge.

Bowles, S. and H. Gintis (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology* **65**(1): 17-28.

Bowles, S. and H. Gintis (2006). Prosocial Emotions. In *The Economy as an Evolving Complex System III* L. Blume and S. N. Durlauf. Eds. Oxford University Press, New York: 339-366.

Bowles, S., H. Gintis and E. Fehr (2003). *Strong Reciprocity May Evolve With or Without Group Selection*, Santa Fe Institute and University of Zurich: 1-8.

Bowles, S. and S.-H. Hwang (2008). Social preferences and public economics: Mechanism design when social preferences depend on incentives. *Journal of Public Economics* **92**(8-9): 1811-1820.

Boyd, R., H. Gintis, S. Bowles and P. J. Richerson (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America* **100**(6): 3531-3535.

Boyd, R. and P. J. Richerson (1988). *Culture and the Evolutionary Process*. University Of Chicago Press, Chicago.

Brams, S. (1994). *Theory of Moves*. Cambridge University Press, London.

Brams, S. and A. D. Taylor (1996). *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press, London.

Brewer, M. B. and L. R. Caporael (1990). Selfish genes vs. selfish people: Sociobiology as origin myth. *Motivation and Emotion* **14**(4): 237-243.

Buchan, N. R., R. T. A. Croson and E. J. Johnson (2002). *Trust and Reciprocity: An International Experiment*. Madison, University of Wisconsin: 1-36.

Buckholtz, J. W., C. L. Asplund, P. E. Dux, D. H. Zald, J. C. Gore, O. D. Jones and R. Marois (2008). The Neural Correlates of Third-Party Punishment. *Neuron* **60**(5): 930-940.

Burnham, T. C. and B. Hare (2007). Engineering Human Cooperation: Does Involuntary Neural Activation Increase Public Goods Contributions? *Human Nature* **18**(2): 88-108.

Burnham, T. C. and D. D. P. Johnson (2005). The Biological and Evolutionary Logic of Human Cooperation. *Analyse & Kritik* **27**(1): 113–135.

Calderon, J. P. and R. Zarama (2006). How Learning Affects the Evolution of Strong Reciprocity. *Adaptive Behaviour* **14**(3): 211-221.

Camerer, C. F. (1997). Progress in Behavioural Game Theory. *The Journal of Economic Perspectives* **11**(4): 167-188.

Camerer, C. F. (1999). Behavioural economics: Reunifying psychology and economics. *Proceedings of the National Academy of Sciences of the United States of America* **96**(19): 10575-10577.

Camerer, C. F. (2003a). Strategizing in the brain. *Science* **300**(5626): 1673(3).

Camerer, C. F. (2003b). *Behavioural Game Theory: Experiments in Strategic Interaction*. Princeton University Press, New Jersey.

Camerer, C. F. (2006). Behavioural Economics. *Advances in Economics and Econometrics* R. Blundell, W. K. Newey and T. Persson. Cambridge University Press, London**:** 181-214.

Camerer, C. F. (2008). Neuroeconomics: Opening the Gray Box. *Neuron* **60**(3): 416-419.

Camerer, C. and R. H. Thaler (1995). Anomalies: Ultimatums, Dictators and Manners. *The Journal of Economic Perspectives* **9**(2): 209-219.

Camerer, C. F., G. Loewenstein and D. Perlec (2004). Neuroeconomics: Why Economics Needs Brains. *Scandinavian Journal of Economics* **106**(3): 555-579.

Camerer, C. F., G. Loewenstein and M. Rabin (2004). *Advances in Behavioural Economics*. Princeton University Press, New Jersey.

Campbell, W. F. (1967). Adam Smith's Theory of Justice, Prudence, and Beneficence. *The American Economic Review* **57**(2): 571-577.

Carraro, C., C. Marchiori, and A. Sgobbi (2005). *Advances in negotiation theory : bargaining, coalitions, and fairness*. Policy Research Working Paper Series. The World Bank.

Carpenter, J., S. Bowles, H. Gintis, and S.-H. Hwang (2008). *Strong Reciprocity and Team Production: Theory and Evidence*. Santa Fe Institute and IZA: 1-31.

Carpenter, J. and P. H. Matthews (2004). *Social Reciprocity*. Bonn, Germany, Institute for the Study of Labor (IZA).

Carter, M. R. and M. Castillo (2002). *The Economic Impacts of Altruism, Trust and Reciprocity: An Experimental Approach to Social Capital*. Wisconsin-Madison Agricultural and Applied Economics Department.

Caruso, R. (2008). Reciprocity in the shadow of threat. *International Review of Economics* **55**(1): 91-111.

Cavalli-Sforza, L. L. and M. W. Feldman (1981). *Cultural Transmission and Evolution*. Princeton University Press, New Jersey.

Chapais, B. (1992). Role of alliances in the social inheritance of rank among female primates. In *Coalitions and Alliances in Humans and Other Animals*. A. H. Harcourt and F. B. M. de Waal. Eds. Oxford University Press, New York**:** 29–60.

Chatterjee, K. and W. Samuelson (1983). Bargaining under Incomplete Information. *Operations Research* **31**(5):835-851.

Chen, K., V. Lakshminarayanan and L.Santos (2005). *The Evolution of Our Preferences: Evidence from Capuchin Monkey Trading Behaviour*, Social Science Research Network.

Chen, S.-H. (2007). Computationally intelligent agents in economics and finance. *Information Sciences* **177**(5): 1153-1168.

Cheney, D. L. (1983). Extra-familial alliances among vervet monkeys. In *Primate Social Relationships*. R. A. Hinde. Blackwell, Oxford**:** 278–286.

Coase, R. H. (1960). The Problem of Social Cost. *The Journal of Law and Economics* **3**(1): 1-69.

Coase, R. H. (1976). Adam Smith's View of Man. *Journal of Law and Economics* **19**(3): 529-546.

Colin, F. C. (2007). Neuroeconomics: Using Neuroscience to Make Economic Predictions. *The Economic Journal* **117**(519): C26-C42.

Colman, A. M. (1999). *Game Theory and Its Applications in the Social and Biological Sciences*. Routledge, New York.

Colman, A. M. (2006). The puzzle of cooperation. *Nature* **440**(7085): 744-745.

Colman, A.M. (2009). *A Dictionary of Psychology (3rd ed)*. Oxford University Press, London.

Cooper, B. and C. Wallace (2004). Group Selection and the Evolution of Altruism. *Oxford Economic Papers* **56**(2): 307-330.

Cory, J. G. A. (2006). A behavioural model of the dual motive approach to behavioural economics and social exchange. *Journal of Socio-Economics* **35**(4): 592-612.

Corsini, R.J. (2002). *The Dictionary of Psychology*. Brunner-Routledge, New York.

Cosmides, L. and J. Tooby (1992). Cognitive Adaptations for Social Exchange. In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. J. Barkow, L. Cosmides and J. Tooby. Eds. Oxford University Press, New York.

Costanza, R., L. Wainger, C. Folke and K.-G. Maler (1993). Modelling complex ecological economic systems. (understanding people and nature). *BioScience* **43**(8): 545-556.

Cramton, P.C. (1992). Strategic Delay in Bargaining with Two-Sided Uncertainty. *Review of Economic Studies* **59**(1):205-225.

Cropsey, J. (1977). *Polity and Economy: An Interpretation of the Principles of Adam Smith.* Greenwood Publication Group, Connecticut.

Cummins, D. D. (1999). Cheater Detection is Modified by Social Rank: The Impact of Dominance on the Evolution of Cognitive Functions. *Evolution and Human Behaviour* **20**(4): 229-248.

Dahlman, C. J. (1979). The Problem of Externality. *Journal of Law and Economics* **22**(1): 141-162.

Daniels, W. M. (1898). The Bearing of the Doctrine of Selection Upon the Social Problem. *International Journal of Ethics* **8**(2): 203-214.

Danielson, P. (2002). Competition among cooperators: Altruism and reciprocity. *Proceedings of the National Academy of Sciences of the United States of America* **99**(Supplement 3): 7237-7242.

Darcet, D. and D. Sornette (2008). Quantitative determination of the level of cooperation in the presence of punishment in three public good experiments. *Journal of Economic Interaction and Coordination* **3**(2): 137-163.

Darwin, C. (1997) [1871]. *The Descent of Man (Great Minds Series)*. Prometheus Books, New York.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, New York.

Debreu, G. (1959). *Theory of Value*. Wiley, New York.

de Quervain, D. J. F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck and E. Fehr (2004). The Neural Basis of Altruistic Punishment. *Science* **305**(5688): 1254-1258.

de Waal, F. B. M. (2000). *Chimpanzee Politics: Power and Sex among Apes*. Johns Hopkins University Press, Baltimore.

Demsetz, H. (1967). Toward a Theory of Property Rights. *The American Economic Review* **57**(2): 347-359.

Dresher, M. (1961). *The Mathematics of Games of Strategy: Theory and Applications.* Prentice-Hall, New Jersey.

Dohmen, T., A. Falk, D. Huffman and U. Sunde (2008). *Homo Reciprocans: Survey Evidence on Behavioural Outcomes*. Maastricht University, Netherlands.

Doughney, J. (2002). *The Poker Machine State: Dilemmas in Ethics, Economics and Governance*. Common Ground Publication, Altona.

Dubreuil, B. (2008). Strong Reciprocity and the Emergence of Large-Scale Societies. *Philosophy of the Social Sciences* **38**(2): 192-210.

Dufwenberg, M., S. Gächter and H. Hennig-Schmidt (2006). *The Framing of Games and the Psychology of Strategic Choice*. Nottingham, Centre for Decision Research and Experimental Economics, University of Nottingham: 1-37.

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games and Economic Behaviour* **47**(2): 268-298.

Durham, W. (1992). *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press, Palo Alto.

Dwyer, J. (2005). Ethics and Economics: Bridging Adam Smith's Theory of Moral Sentiments and Wealth of Nations. *The Journal of British Studies* **44**(4): 662-687.

Eckel, C. C. (2004). Vernon Smith: economics as a laboratory science. *Journal of Socio-Economics* **33**(1): 15-28.

Eckel, C. C. and R. K. Wilson (1998). Reciprocal Fairness and Social Signaling: Experiments with Limited Reputations. *American Economic Association Meeting*. New York.

Edgeworth, F. Y. (1881). *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. Kegan Paul, London.

Ehrlich, P. R. and S. A. Levin (2005). The Evolution of Norms. *PLoS Biology* **3**(6): e194.

Eldakar, O. T., D. L. Farrell and D.S. Wilson (2007). Selfish punishment: Altruism can be maintained by competition among cheaters. *Journal of Theoretical Biology* **249**(2): 198-205.

Eldakar, O. T. and D. S. Wilson (2008). Selfishness as second-order altruism. *Proceedings of the National Academy of Sciences* **105**(19): 6982-6986.

Ellingsen, T. (1997). The Evolution of Bargaining Behavior. *The Quarterly Journal of Economics* **112** (1): 581–602.

Elster, J. (1989). Social Norms and Economic Theory. *The Journal of Economic Perspectives* **3**(4): 99-117.

Eriksson, A. and K. Lindgren (2005). Cooperation driven by mutations in multi-person Prisoners' Dilemma. *Journal of Theoretical Biology* **232**(3): 399-409.

Evensky, J. (2005). Adam Smith's Theory of Moral Sentiments: On Morals and Why They Matter to a Liberal Society of Free People and Free Markets. *Journal of Economic Perspectives* **19**(3): 109–130.

Falk, A. and U. Fischbacher (2006). A theory of reciprocity. *Games and Economic Behaviour* **54**(2): 293-315.

Fehr, E. and C. F. Camerer (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences* **11**(10): 419-427.

Fehr, E. and A. Falk (1999). Wage Rigidity in a Competitive Incomplete Contract Market. *The Journal of Political Economy* **107**(1): 106-134.

Fehr, E. and U. Fischbacher (2002). Why Social Preferences Matter - The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives. *The Economic Journal* **112**(478): C1-C33.

Fehr, E. and U. Fischbacher (2003). The nature of human altruism. *Nature* **425**(6960): 785-791.

Fehr, E. and U. Fischbacher (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences* **8**(4): 185-190.

Fehr, E. and U. Fischbacher (2004b). Third-party punishment and social norms. *Evolution and Human Behaviour* **25**(2): 63-87.

Fehr, E. and U. Fischbacher (2005). Human Altruism – Proximate Patterns and Evolutionary Origins. *Analyse & Kritik* **27**(1): 6–47.

Fehr, E., U. Fischbacher and S. Gächter (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* **13**(1): 1-25.

Fehr, E. and S. Gächter (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review* **90**(4): 980-994.

Fehr, E. and S. Gächter (2002). Altruistic punishment in humans. *Nature* **415**(6868): 137(4).

Fehr, E., S. Gächter and G. Kirchsteiger (1996). Reciprocal Fairness and Noncompensating Wage Differentials. *Journal of Institutional and Theoretical Economics* **152**(4): 608-640.

Fehr, E. and H. Gintis (2007). Human Motivation and Social Cooperation: Experimental and Analytical Foundations. *Annual Review of Sociology* **33**(1): 43.

Fehr, E. and J. Henrich (2002). Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism. In *Genetic and Cultural Evolution of Cooperation*. P. Hammerstein. Eds. MIT Press, Cambridge**:** 55-76.

Fehr, E., G. Kirchsteiger, and A.Riedl (1993). Does Fairness Prevent Market Clearing? An Experimental Investigation. *The Quarterly Journal of Economics* **108**(2): 437-459.

Fehr, E., G. Kirchsteiger and A. Riedl (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review* **42**(1): 1-34.

Fehr, E. and B. Rockenbach (2003). Detrimental effects of sanctions on human altruism. *Nature* **422**(6928): 137(4).

Fehr, E. and B. Rockenbach (2004). Human altruism: economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology* **14**(6): 784-790.

Fehr, E. and K. M. Schmidt (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* **114**(3): 817-868.

Fehr, E. and K. M. Schmidt (2006). The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories. In *Handbook of the Economics of Giving, Altruism and Reciprocity*. S. C. Kolm and J. M. Ythier. Eds. North Holland, Paris.

Fehr, E. and F. Schneider (2007). Implicit Reputation Cues and Strong Reciprocity. In *10th IZA European Summer School in Labor Economics*. A. R. Cardoso and K. F. Zimmermann. Eds. (2007) Ammersee Lake, Germany.

Fehr, E., G. Simon and G. Kirchsteiger (1997). Reciprocity as a Contract Enforcement Device: Experimental Evidence. *Econometrica* **65**(4): 833-860.

Fishman, M. A., A. Lotem and L. Stone (2001). Heterogeneity Stabilizes Reciprocal Altruism Interactions. *Journal of Theoretical Biology* **209**(1): 87-95.

Flood, M.M. (1952). Some experimental games. Research memorandum RM-789. RAND Corporation, Santa Monica, CA.

Force, P. (1997). Self-Love, Identification, and the Origin of Political Economy. *Yale French Studies*(92): 46-64.

Force, P. (2003). *Self-interest Before Adam Smith: A Genealogy of Economic Science*. Cambridge University Press, Cambridge.

Frantz, R. (2000). Intuitive elements in Adam Smith. *Journal of Socio-Economics* **29**(1): 1-19.

Frith, C. D. and U. Frith (2008). Implicit and Explicit Processes in Social Cognition. *Neuron* **60**(3): 503-510.

Fudenberg, D. and J. Tirole (1991). *Game Theory*. MIT Press, Cambridge.

Gächter, S. and A. Falk, 2001. Reputation or Reciprocity? An Experimental Investigation, *CESifo Working Paper Series 496*, CESifo Group Munich.

Gächter, S. and E. Fehr (2004). Fairness and Retaliation: The Economics of Reciprocity. In *Advances in Behavioural Economics* C. F. Camerer, G. Loewenstein and M. Rabin. Eds. Princeton University Press, Princeton**:** 510-532.

Gächter, S. and B. Herrmann (2006). Human cooperation from an economic perspective. In *Cooperation in Primates and Humans: Mechanisms and Evolution* P. M. Kappeler and C. P. v. Schaik. Eds. Springer, Berlin: 279-301.

Geanakoplos, J., D. Pearce and E. Stacchetti (1989). Psychological games and sequential rationality. *Games and Economic Behavior* **1**:60-79.

Gillies, A. S. and M. L. Rigdon (2008). *Epistemic Conditions and Social Preferences in Trust Games*. University Library of Munich, Germany.

Gintis, H. (2000a). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology* **206**(2): 169-179.

Gintis, H. (2000b). Beyond Homo economicus: evidence from experimental economics. *Ecological Economics* **35**(3): 311-322.

Gintis, H. (2000c). *Game theory evolving*. Princeton University Press, New Jersey.

Gintis, H. (2005). Behavioural Game Theory and Contemporary Economic Theory. *Analyse & Kritik* **27**(1): 48-72.

Gintis, H. (2007). Modelling Cooperation with Self-regarding Agents. Santa Fe Institute, Santa Fe, New Mexico.

Gintis, H. (2009). *Game Theory Evolving: A Problem-Centered introduction to Modelling Strategic Interaction*. Princeton University Press, Princeton.

Gintis, H., S. Bowles, R. Boyd and E. Fehr (2003). Explaining altruistic behaviour in humans. *Evolution and Human Behaviour* **24**(3): 153-172.

Gintis, H., S. Bowles, R. Boyd and E. Fehr (2005). *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life*. MIT Press, Cambridge.

Gintis, H., J. Henrich, R. Boyd and E. Fehr (2008). Strong Reciprocity and the Roots of Human Morality. *Social Justice Research* **21**(2): 241-253.

Glimcher, P. W., C. Camerer, C.F., E. Fehr, and R. A. Poldrack. Eds. (2008). *Neuroeconomics: Decision Making and the Brain*. Academic Press, New York.

Gosling, T., N. Jin and E. P. K Tsang (2005). Population based incremental learning with guided mutation versus genetic algorithms: iterated prisoners dilemma. A. Auger. Eds. *Evolutionary Computation, 2005. The 2005 IEEE Congress on*. **1**: 958-965 Vol.1.

Gotts, N. M., J. G. Polhill and A. N. R. Law (2003). Agent-Based Simulation in the Study of Social Dilemmas. *Artificial Intelligence Review* **19**(1): 3-92.

Gramm, W. S. (1989). The Selective Interpretation of Adam Smith. In *The Methodology of Economic Thought*. M. R. Tool and W. J. Samuels. Eds. Transaction Publishers, Edison**:** 590

Gray, J. R. and T. S. Braver (2002). Cognitive control in altruism and self-control: A social cognitive neuroscience perspective. *Behavioural and Brain Sciences* **25**(02): 260-260.

Guth, W. (1995). An evolutionary approach to explaining cooperative behaviour by reciprocal incentives. *International Journal of Game Theory* **24**: 323-344.

Guth, W. and R. Tietz (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology* **11**: 417—449.

Guth, W. and M. E. Yaari (1992). Explaining Reciprocal Behaviour in Simple Strategic Games: An Evolutionary Approach. In *Explaining Process and Change: Approaches to Evolutionary Economics*. University of Michigan Press, Ann Arbor: 23-34.

Hagen, E. H. and P. Hammerstein (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology* **69**(3): 339-348.

Hamilton W.D. (1964). The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology* **7**: 1-16 and 17-52.

Hamilton W.D. (1975). Innate social aptitudes of man: an approach from evolutionary genetics. in R. Fox (ed.). *Biosocial Anthropology*. Malaby Press, London, 133-53.

Hansson, I. and C. Stuart (1992). Socialization and altruism. *Journal of Evolutionary Economics* **2**(4): 301-312.

Hardin, G. (1968). The Tragedy of the Commons. *Science* **162**(3859): 1243-1248.

Harsyani, J.C. (1966) A general theory of rational behavior in game situations. *Econometrica* **34** (3): 613-634.

Harsyani, J.C. (1967) Games with incomplete information played by 'Bayesian' disputants, part I: The basic model. *Management Science* **14**(3): 159-182.

Harsyani, J.C. (1968a) Games with incomplete information played by 'Bayesian' disputants, part II: Bayesian equilibrium points. *Management Science* **14**(5): 320-334.

Harsyani, J.C. (1968b) Games with incomplete information played by 'Bayesian' disputants, part III: The basic probability distribution of the game. *Management Science* **14** (7): 486-502.

Hauert, C., S. De Monte and K. Sigmund (2002). Replicator Dynamics for Optional Public Good Games. *Journal of Theoretical Biology* **218**(2): 187-194.

Hechter, M. and K. D. Opp. Eds. (2001). *Social Norms*. Russell Sage Foundation, New York.

Heilbroner, R. L. (1973). The Paradox of Progress: Decline and Decay in The Wealth of Nations. *Journal of the History of Ideas* **34**(2): 243-262.

Henrich, J., R. Boyd, S. Bowles, S. Camerer, E. Fehr, H. Gintis, and R. McElreach (2001). In Search of Homo Economicus: Behavioural Experiments in 15 Small-Scale Societies. *The American Economic Review* **91**(2): 73-78.

Hill, J. (1984). Human altruism and sociocultural fitness. *Journal of Social and Biological Structures* **7**: 17-35.

Hirshleifer, D. and E. Rasmusen (1989). Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behaviour & Organization* **12**: 87-106.

Hirshleifer, J. (1977). Economics from a Biological Viewpoint *Journal of Law & Economics* **20**(1): 1-52.

Hirshleifer, J. (1978). Competition, Cooperation, and Conflict in Economics and Biology. *The American Economic Review* **68**(2): 238-243.

Hirshleifer, J. (1985). The Expanding Domain of Economics. *The American Economic Review* **75**(6): 53-68.

Hobbes, T. (1968). *The Leviathan*. Penguin Books, London.

Hoffman, E., K. A. McCabe and V. L. Smith (1998). Behavioural foundations of reciprocity: experimental economics and evolutionary psychology. *Economic Inquiry* **36**(3): 335-353.

Holland, J. H. and J. H. Miller (1991). Artificial Adaptive Agents in Economic Theory. *The American Economic Review* **81**(2): 365-370.

Holmstrom, B. and R. B. Myerson (1983). Efficient and Durable Decision Rules with Incomplete Information. *Econometrica* **51**(6):1799-1819.

Holt, C. A. and S. K. Laury (1997). *Voluntary Provision of a Public Good*. University of Virginia, Charlottesville**:** 1-10.

Homans, G. C. (1958). Social Behaviour as Exchange. *The American Journal of Sociology* **63**(6): 597-606.

Hopmann, P.T. (1995). Two Paradigms of Negotiation: Bargaining and Problem Solving. *The Annals of the American Academy of Political and Social Science* **542**(1): 24-47.

Huberman, B. A. and N. S. Glance (1993). Evolutionary games and computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **90**(16): 7716-7718.

Huck, S. and J. Oechssler (1999). The Indirect Evolutionary Approach to Explaining Fair Allocations. *Games and Economic Behaviour* **28**(1): 13-24.

Kahana, N. (2005). On the Surge of Altruism. *Journal of Population Economics* **18**(2): 261-266.

Kahneman, D. (2003a). A Psychological Perspective on Economics. *The American Economic Review* **93**(2): 162-168.

Kahneman, D. (2003b). Maps of Bounded Rationality: Psychology for Behavioural Economics. *The American Economic Review* **93**(5): 1449-1475.

Kahneman, D., J. L. Knetsch and R. H. Thaler (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American Economic Review* **76**(4): 728-741.

Kahneman, D., J. L. Knetsch and R. H. Thaler (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. *The Journal of Political Economy* **98**(6): 1325-1348.

Kahneman, D. and A. Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**(2): 263-291.

Kangas, O. E. (1997). Self-interest and the common good: The impact of norms. *Journal of Socio-Economics* **26**(5): 475.

Kaplow, L. and S. Shavell (2007). Moral Rules, the Moral Sentiments, and Behaviour: Toward a Theory of an Optimal Moral System. *The Journal of Political Economy* **115**(3): 494-514.

Kennett, D. A. (1980a). Altruism and Economic Behaviour, I Developments in the Theory of Public and Private Redistribution. *American Journal of Economics and Sociology* **39**(2): 183-198.

Kennett, D. A. (1980b). Altruism and Economic Behaviour: II Private Charity and Public Policy. *American Journal of Economics and Sociology* **39**(4): 337-352.

Ketelaar, T., B. Preston, D. Russell, M. Davis and G. Strosser (2007). EMOTLAB: Software for studying emotional signaling in economic bargaining games. *Behaviour Research Methods* **39**(4): 959-972.

Khalil, E. L. (2001). Adam Smith and Three Theories of Altruism. *Recherches économiques de Louvain* **67**(4): 421-435.

Khalil, E. L. (2004). Is a group better off with more altruists? Not necessarily. *Journal of Economic Behaviour & Organization* **53**(1): 89-92.

Khalil, E. L. (2004a). What is altruism? *Journal of Economic Psychology* **25**(1): 97-123.

Khalil, E. L. (2004b). What is altruism? A reply to critics. *Journal of Economic Psychology* **25**(1): 141-143.

Khalil, E. L. (2007). *The Mirror-Neuron Paradox: How Far is Sympathy from Compassion, Indulgence, and Adulation?* Department of Economics, Monash University (Clayton), Australia**:** 1-68.

Kim, S.Y. and C. Taber (2004). A Cognitive/Affective Model of Strategic Behaviour - two-person Repeated Prisoners' Dilemma Game. *Proceedings of the Sixth International Conference on Cognitive Modelling*. Lawrence Earlbaum, New Jersey**:** 360-361.

Kitcher, P. (1998). Psychological Altruism, Evolutionary Origins, and Moral Rules. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* **89**(2/3): 283-316.

Klein, R. A. (2008). *The Neurobiology of Altruistic Punishment*. Biological Explanations of Behaviour: Philosophical Perspectives Hannover, Germany.

Klein, S. (2003). The Natural Roots of Capitalism and Its Virtues and Values. *Journal of Business Ethics* **45**(4): 387-401.

Kockesen, L., E. Ok and R. Sethi (2000). Evolution of interdependent preferences in aggregative games. *Games and Economic Behavior* **31**: 303-310

Kolm, S.-C. (1983). Altruism and Efficiency. *Ethics* **94**(1): 18-65.

Kolm, S.-C., L.-A. Gérard-Varet and J. Mercier Ythier. Eds. (2006). *Handbook on the Economics of Giving, Reciprocity and Altruism*. Elsevier, New York.

Komorita, S. S., C. D. Parks and L. G. Hulbert (1992). Reciprocity and the induction of cooperation in social dilemmas. *Journal of Personality and Social Psychology* **62**(4): 607-617.

Koslowski, P. Ed. (1999). *The Theory of Evolution in Biological and Economic Theory*. Springer, Cambridge.

Kramer, R. M., E. Newton and P. L. Pommerenke (1993). Self-enhancement biases and negotiator judgment: Effects of self-esteem and mood. *Organizational Behavior and Human Decision Processes* **56**:110-133.

Krajbich, I., R. Adolphs, D. Tranel, N. Denburg and C. F. Camerer (2009). Economic Games Quantify Diminished Sense of Guilt in Patients with Damage to the Prefrontal Cortex. *Journal of Neuroscience* **29**(7): 2188-2192.

Krebs, D. (1987). The challenge of altruism in biology and psychology. In *Sociobiology and Psychology: Ideas, Issues, and Applications* M. Smith, C. Crawford and D. Krebs. Eds. Lawrence Erlbaum, Philadelphia.

Kreps, D.M. and R. Wilson (1982). Sequential Equilibria. *Econometrica* **50**(4): 863-894.

Krueger, F., J. Grafman and K. McCabe (2008). Neural correlates of economic game playing. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1511): 3859-3874.

Krueger, F., K. McCabe, J. Moll, N. Kriegeskorte, R. Zahn, M. Strenziok, A. Heinecke and J. Grafman (2005). *Neural Correlates of Conditional and Unconditional Trust in Two-Person Reciprocal Exchange*. Cognitive Neuroscience Section, NINDS, National Institutes of Health, Bethesda, Maryland, USA: 1-20.

Lamb, R. B. (1974). Adam Smith's System: Sympathy not Self-Interest. *Journal of the History of Ideas* **35**(4): 671-682.

Leist, A. (2005). Social Relations Instead of Altruistic Punishment. *Analyse & Kritik* **27**(1): 158-171.

Levine, D. K. (1998). Modelling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* **1**: 593–622.

Levine, D. S. (2006). Neural modelling of the dual motive theory of economics. *Journal of Socio-Economics* **35**(4): 613-625.

Lewis, J. S. (2006). *The Function of Free Riders: Toward a Solution to the Problem of Collective Action*. Bowling Green State University, Graduate College. Ohio, USA. **Doctor of Philosophy:** 192.

Liebrand, W. B. G. (1986). The Ubiquity of Social Values in Social Dilemmas. In *Experimental Social Dilemmas* H. A. M. Wilke, D. M. Messick and C. G. Rutte. Eds. Peter Lang Publishing, Bern: 113-134.

Liu, X., D. K. Powell, H. Wang, B. T. Gold, C. R. Corbly (2007). Functional Dissociation in Frontal and Striatal Areas for Processing of Positive and Negative Reward Information. *Journal of Neuroscience* **27**(17): 4587-4597.

Lloyd, B. (2007). The Commons revisited: The tragedy continues. *Energy Policy* **35**(11): 5806-5818.

Loewenstein, G. F., L. Thompson and M. H. Bazerman (1989). Social Utility and Decision Making in Interpersonal Contexts. *Journal of Personality and Social Psychology* **57**(3):426-441.

Luca, A. and F. Leonardo (2006). Transaction Costs and the Robustness of the Coase Theorem. *Economic Journal* **116**(508): 223-245.

Luce, R.D. and H. Raiffa (1957). *Game and Decisions: Introduction and critical survey*. John Wiley, New York.

Lynne, G. D. (2006). Toward a dual motive metaeconomic theory. *Journal of Socio-Economics* **35**(4): 634-651.

Macfie, A. L. (1959). Adam Smith's Moral Sentiments as Foundation for His Wealth of Nations. *Oxford Economic Papers* **11**(3): 209-228.

Maitland, I. (2002). The Human Face of Self-Interest. *Journal of Business Ethics* **38**(1): 3-17.

Mancur, O. (1971). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge.

Manhart, K. (2007). *Cooperation in 2- and N-Person Prisoners' Dilemma Games: A Simulation Study*. University Bern, Switzerland**:** 1-14.

Manski, C. F. (2000). Economic Analysis of Social Interactions. *The Journal of Economic Perspectives* **14**(3): 115-136.

Margolis, H. (1984). *Selfishness, Altruism, and Rationality* University Of Chicago Press, Chicago.

Marlowe, F. W., J. C. Berbesque, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, J. Ensminger, M. Gurven, E. Gwako, J. Henrich, N. Henrich, C. Lesorogol, R. McElreath and D. Tracer (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences* **275**(1634): 587-592.

Masclet, D. and M.-C. Villeval (2006). Punishment, Inequality and Emotions, Groupe d'Analyse et de Thiorie Economique (GATE), Centre national de la recherche scientifique (CNRS), Universiti Lyon 2, Ecole Normale Superieure, France.

Marynard Smith, J. (1982). *Evolution and the Theory of Games.* Cambridge University Press, Cambridge.

Mayr, U., W. T. Harbaugh and D. and Tankersley (2008). Neuroeconomics of Charitable Giving and Philanthropy. In *Neuroeconomics: Decision Making and the Brain*. P. W. Glimcher, C. Camerer, R. A. Poldrack and E. Fehr. Eds. Academic Press, New York.

McCabe, K. A., S. J. Rassenti and V. L. Smith (1996). Game theory and reciprocity in some extensive form experimental games. *Proceedings of the National Academy of Sciences of the United States of America* **93**(23): 13421-13428.

McCabe, K. A., M. L. Rigdon and V. L. Smith (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behaviour & Organization* **52**(2): 267-275.

Mendes, R. V. (2004). Network Dependence of Strong Reciprocity. *Advances in Complex Systems* **7**(3/4): 357-368.

Messick, D. M. (1992). Equality as a decision heuristic. In *Psychological Issues in Distributive Justice*. B. Mellers. Eds. Springer-Verlag, New York.

Messick, D. M., and T. Schell (1992). *Evidence for an equality heuristic in social decision making*. Working Paper No. 83. Dispute Resolution Research Center, Kellogg Graduate School of Management, Northwestern University, Evanston.

Michael, R. T. and G. S. Becker (1973). On the New Theory of Consumer Behaviour. *Swedish Journal of Economics* **75**(4): 378.

Miller, T. C. (1993). The Duality of Human Nature. *Politics and the Life Sciences* **12**(2): 221-241.

Montague, P. R. and T. Lohrenz (2007). To Detect and Correct: Norm Violations and Their Enforcement. *Neuron* **56**(1): 14-18

Montes, L. (2003). Das Adam Smith Problem : Its Origins, the Stages of the Current Debate, and One Implication for our Understanding of Sympathy. *Journal of the History of Economic Thought* **25**(1): 63 - 90.

Morrow, G. R. (1969). *The Ethical and Economic Theories of Adam Smith*. Augustus M. Kelley Publishers, New York.

Mueller, D. C. (1986). Rational Egoism versus Adaptive Egoism as Fundamental Postulate for a Descriptive Theory of Human Behaviour. *Public Choice* **51**(1): 3-23.

Mullainathan, S. and R. H. Thaler (2000). Behavioural Economics. *MIT Dept. of Economics Working Paper No. 00-27,* Cambridge.

Muller, J. Z. (1995). *Adam Smith in His Time and Ours: Designing the Decent Society*. Princeton University Press, New Jersey.

Musgrave, R.A. (1959). *The Theory of Public Finance: A Study in Public Economy*. McGraw-Hill, New York.

My, K. B. and B. Chalvignac (2007). Learning and coping in repeated collective-good games. *Conference of the French Economic Association on Behavioural Economics and Experiments* Lyon, Journée d'Economie Expérimentale, France.

Myerson, R. B. (1977) Graphs and Cooperation in Games. *Mathematics of Operations Research* **2**: 225–229.

Myerson, R.B. (1979). Incentive compatibility and the bargaining problem. *Econometrica* **47**: 61-73.

Myerson, R.B. (1984). Two-Person Bargaining Problems with Incomplete Information. Econometrica 52(2): 461-487.

Myerson, R.B. (1985). Negotiation in Games: A Theoretical Overview. Discussion Papers. Center for Mathematical Studies in Economics and Management Science. Northwestern University, Evanston.

Nakamaru, M. and Y. Iwasa (2006). The coevolution of altruism and punishment: Role of the selfish punisher. *Journal of Theoretical Biology* **240**(3): 475-488.

Nash, J.F. (1950) The Bargaining Problem. *Econometrica* **18**: 155–162.

Nash J.F. (1951) Non-cooperative games. *Annals of Mathematics* **54**: 286-295.

Nash, J.F. (1953). Two-Person Cooperative Games. *Econometrica* **21**(1): 128-140.

Neumärker, B. (2007). Neuroeconomics and the Economic Logic of Behaviour. *Analyse & Kritik* **29**: 60-85.

Neyman, A. (1985). Bounded complexity justifies cooperation in the finitely repeated prisoners' dilemma. *Economics Letters* **19**(3): 227-229.

Nieli, R. (1986). Spheres of Intimacy and the Adam Smith Problem. *Journal of the History of Ideas* **47**(4): 611-624.

Neale, M.A and G.B. Northcroft (1991). Behavioral Negotiation Theory: A framework of Dyadic Bargaining. *Research in Organizational Behavior*. **13**:147-190.

Nowak, M. A. and K. Sigmund (2005). Evolution of indirect reciprocity. *Nature* **437**(7063): 1291(8).

Nussbaum, M.C. (1997), Flawed Foundations: The Philosophical Critique of (A Particular Type of) Economics. *University of Chicago Law Review*. **64**(4): 1197-1214.

Ochs, J. and A. E. Roth (1989). An Experimental Study of Sequential Bargaining. *American Economic Review* **79**: 355–384.

Odling-Smee, F. J., K. N. Laland and M. W. Feldman (2003). *Niche Construction: The Neglected Process in Evolution*. Princeton University Press, New Jersey.

O'Doherty, J., M. L. Kringelbach, E. T. Rolls, J. Hornak and C. Andrews (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience* **4**(1): 95.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review* **46**(8): 1423-1437.

Olson, M. (1979). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge.

Osborne, M. J., A. Rubenstein (1990). *Bargaining and Markets*. Academic Press, London.

Ostmann, A. and H. Meinhardt (2007). Toward an Analysis of Cooperation and Fairness That Includes Concepts of Cooperative Game Theory. In *New Issues and Paradigms in Research on Social Dilemmas*. A. Biel, D. Eek, T. Gärling and M. Gustafson. Eds. Springer, New York: 230-251.

Ostrom, E., R. Gardner and J. Walker (1994). *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor.

Panchanathan, K. and R. Boyd (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**(7016): 499(4).

Peacock, M. S. (2007). The Conceptual Construction of Altruism: Ernst Fehr's Experimental Approach to Human Conduct. *Philosophy of the Social Sciences* **37**(1): 3-23.

Peressini, A. (1993). Generalizing Evolutionary Altruism. *Philosophy of Science* **60**(4): 568-586.

Peters-Fransen, I. (2001). The canon in the history of the Adam Smith problem. In *Reflections on the Classical Canon in Economics: Essays in Honor of Samuel Hollander*. E. L. Forget, S. Hollander and S. Peart. Eds. Routledge, New York**:** 168-184.

Piliavin, J. A. (1981). *Emergency Intervention*. Academic Press, St Louis.

Piliavin, J. A. and H.-W. Charng (1990). Altruism: A Review of Recent Theory and Research. *Annual Review of Sociology* **16**: 27-65.

Pingle, M. Deliberation Cost as a Foundation for Behavioural Economics. In *Handbook of Contemporary Behavioural Economics: Foundations and Developments*. M. Altman. Eds. M.E. Sharpe, Armonk, New York**:** 340-378.

Poncela, J., J. Gomez-Gardenes, L. M. Floría and Y Moreno (2007). Robustness of cooperation in the evolutionary Prisoners' Dilemma on complex networks. *New Journal of Physics* **9**(6): 184-184.

Posner, E. A. (2002). *Law and Social Norms.* Harvard University Press, Cambridge.

Price, M. E., L. Cosmides and J. Tooby (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behaviour* **23**(3): 203-231.

Pruitt, D. and J. Z. Rubin (1986). *Social Conflict: Escalation, Stalemate and Settlement.* McGraw Hill, New York.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review* **83**(5): 1281-1302.

Rachlin, H. (2002). Altruism and selfishness. *Behavioural and Brain Sciences* **25**(02): 239-250.

Raphael, D.D. and A.L. Macfie. (1976). *Introduction. The Theory of Moral Sentiments*. Clarendon Press. Oxford.

Rastetter, E. B., A. W. King, B. J. Cosby, G. M. Hornberger, R. V. O'Neill and J. E. Hobbie (1992). Aggregating fine-scale ecological knowledge to model coarser-scale attributes of ecosystems. *Ecological Applications* **2**:55-70.

Reeve, S. (1906). *The cost of competition: an effort at the understanding of familiar facts.* McClure, Phillips & Co, New York.

Rick, S. and G. F. Loewenstein (2007). *The Role of Emotion in Economic Behaviour*. University of Pennsylvania - The Wharton School, Philadelphia.

Rigdon, M. L., K. A. McCabe and V. L. Smith (2007). Sustaining Cooperation in Trust Games. *The Economic Journal* **117**(522): 991-1007.

Rilling, J. K., D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts (2002). A Neural Basis for Social Cooperation *Neuron* **35**(2): 395-405.

Rilling, J. K., B. King-Casas and A. G. Sanfey (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology* **18**(2): 159-165.

Riolo, R. L., M. D. Cohen and R. Axelrod (2001). Evolution of cooperation without reciprocity. *Nature* **414**(6862): 441(3).

Robbins, L. (1952). *The Theory of Economic Policy in English Classical Political Economy*. Macmillan, London.

Robinson, J. (1962). *Economic Philosophy*. Penguin Books, Harmondsworth, UK.

Robson, A. J. (2001). The Biological Basis of Economic Behaviour. *Journal of Economic Literature* **39**(1): 11-33.

Rossi, A. and M. Warglien (2000). *An Experimental Investigation of Fairness and Reciprocal Behaviour in a Simple Principal-Multiagent Relationship*. Department of Computer and Management Sciences, University of Trento, Italy.

Roth, A.E. (1983) Towards a theory of bargaining: An experimental study in economics. *Science* **220**: 687-690.

Roth, A.E. (1985) A note on risk aversion in a perfect equilibrium model of bargaining. *Econometrica* **53**: 207-211.

Roth, A. E. (2002). The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica* **70**(4): 1341-1378.

Roth, A.E. and F. Schoumaker (1983) Expectations and Reputations in bargaining: An experimental study. *American Economic Review* **73**(3): 362-372.

Roth, A.E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991). Bargaining and market behavior in Jerusalem, Ljubiljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review* **81**: 1068-1095.

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica* **50**(1): 97-109.

Rubinstein, A. (1985). A Bargaining Model with Incomplete Information about Time Preferences. *Econometrica* **53**(5): 1151-1172.

Rustichini, A. (2005). Neuroeconomics: Present and future. *Games and Economic Behaviour* **52**(2): 201-212.

Rydz, D. (2005). Altruism: Dare We Believe? *University of Alberta Health Sciences Journal* **2**(2): 30-32.

Sanfey, A. G. (2007). Social Decision-Making: Insights from Game Theory and Neuroscience. *Science* **318**(5850): 598-602.

Sanfey, A. G., J. K. Rilling, J. A. Aronson, L. E. Nystrom and J. D. Cohen (2003). The neural basis of economic decision-making in the ultimatum game *Science* **300**(5626): 1755(4).

Santos, L. R. and K. M. Chen (2008). The Evolution of Rational and Irrational Economic Behaviour: Evidence and Insight from a Non-Human Primate Species. In

Neuroeconomics: Decision Making and the Brain. P. W. Glimcher, C. Camerer, R. A. Poldrack and E. Fehr. Eds. Academic Press, New York**:** 81-92.

Satterthwaite, M. and S. R. Williams (1989). The Rate of Convergence to Efficiency in the Buyers' Bid Double Auction as the Market Becomes Large. *Review of Economic Studies* **56**(4):477-498.

Schelling, T. C. (1968). Game Theory and the Study of Ethical Systems. *The Journal of Conflict Resolution* **12**(1): 34-44.

Schenk, R. E. (1987). Altruism as a Source of Self-Interested Behaviour. *Public Choice* **53**(2): 187-192.

Scherbaum, S., M. Dshemuchadse and A. Kalis (2008). Making decisions with a continuous mind. *Cognitive, Affective, & Behavioural Neuroscience* **8**(4): 454-474.

Searle, J. (1990). Collective Intentions and Actions. In *Intentions in Communications*. P. Cohen, J. Morgan and M. E. Pollack. Eds. MIT Press, Cambridge.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* **4**:25-55.

Sen, A. K. (1977). Rational Fools: A Critique of the Behavioural Foundations of Economic Theory. *Philosophy and Public Affairs* **6**(4): 317-344.

Sen, A. K. (1993). Does Business Ethics Make Economic Sense? *Business Ethics Quarterly* **3**(1): 45-54.

Sen, A. K. (2009). *The Idea of Justice*. Harvard University Press, Cambridge.

Sesardic, N. (1995). Recent Work on Human Altruism and Evolution. *Ethics* **106**(1): 128-157.

Sethi, R. and E. Somanathan (2004). What can we learn from cultural group selection and co-evolutionary models? *Journal of Economic Behaviour & Organization* **53**(1): 105-108.

Sethi, R. and E. Somanathan (2005). Norm Compliance and Strong Reciprocity. In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. H. Gintis, S. Bowles, R. T. Boyd and E. Fehr. Eds. MIT Press, Cambridge**:** 229-250.

Shapley, L. S. (1953). A Value for n-Person Games. In *Contributions to the Theory of Games II (Annals of Mathematics Studies)*. H. W. Kuhn and A. W. Tucker. Eds. Princeton University Press, Princeton: 307–317

Shapley, L. S. (1977). *A Comparison of Power Indices and a Nonsymmetric Generalization*, P-5872. Rand Corporation, Santa Monica.

Shapley, L. S. and M. Shubik (1954). A Method for Evaluating the Distribution of Power in a Committee System. *American Political Science Review* **48**: 787–792

Shermer, M. (2007). *The Mind of the Market: Compassionate Apes, Competitive Humans, and Other Tales from Evolutionary Economics*. Macmillan, New York.

Shinada, M., T. Yamagishi and Y. Yamamoto (2004). False friends are worse than bitter enemies: Altruistic punishment of in-group members. *Evolution and Human Behaviour* **25**(6): 379-393.

Shogren, J. F. and L. O. Taylor (2008). On Behavioural-Environmental Economics. *Review of Environmental Economics and Policy* **2**(1): 26-44.

Sigmund, K., C. Hauert and M. A. Nowak (2001). Reward and punishment. *Proceedings of the National Academy of Sciences of the United States of America* **98**(19): 10757-10762.

Silk, J.B. (2005). The evolution of cooperation in primate groups. In: *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life*. H. Gintis, S. Bowles, R. Boyd, and E. Fehr. Eds. MIT Press, Cambridge.

Simmons, R. L. and S. K. Marine (2002). *Gift of Life: The Effect of Organ Transplantation on Individual, Family, and Societal Dynamics*. Transaction Publishers, Edison.

Simon, H. (1992). Altruism and Economics. *Eastern Economic Journal*. **18** (1):73-83.

Singer, A. E. (1994). Strategy as Moral Philosophy. *Strategic Management Journal* **15**(3): 191-213.

Singer, T. and E. Fehr (2005). The Neuroeconomics of Mind Reading and Empathy. *The American Economic Review* **95**(2): 340-345.

Smirnov, O., C. T. Dawes, J. H. Fowler, T. Johnson and R. McElreath (2007). The Behavioural Logic of Collective Action: Partisans Cooperate and Punish More than Non-Partisans. *Political Psychology*: 1-38.

Smith, A. (1759) [1976]. *The Theory of Moral Sentiments*. D.D. Raphael and A.L. Macfie. Eds. Clarendon Press, Oxford.

Smith, A. (1776) [1976]. *An Inquiry into the Nature and Causes of the Wealth of Nations*. E. Cannan. Eds. University Of Chicago Press, Chicago.

Smith, A. (1790) [2000]. *The Theory of Moral Sentiments.* Prometheus Books, New York.

Smith, V. L. (1998). The Two Faces of Adam Smith. *Southern Economic Journal* **65**(1): 2-19.

Smith, V. L. (2005). Behavioural economics research and the foundations of economics. *Journal of Socio-Economics* **34**(2): 135-150.

Soanes, C. and A. Stevenson (2005). *Oxford Dictionary of English*. Oxford University Press, Oxford.

Spiekermann, K. (2008). *Norms and Games: Realistic Moral Theory and the Dynamic Analysis of Cooperation.* Political Science. London, London School of Economics and Political Science. **Doctor of Philosophy:** 182.

Spitzer, M., U. Fischbacher, B. Herrnberger, G. Gron and E. Fehr (2007). The Neural Signature of Social Norm Compliance. *Neuron* **56**(1): 185-196

Stephens, C. (2005). Strong Reciprocity and the Comparative Method. *Analyse & Kritik* **27**(1): 97-105.

Stevens, J. R. and M. D. Hauser (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences* **8**(2): 60-65.

Sugden, R. (1984). Reciprocity: The Supply of Public Goods Through Voluntary Contributions. *The Economic Journal* **94**(376): 772-787.

Suttle, B. B. (1987). The Passion of Self-interest: The Development of the Idea and Its Changing Status. *American Journal of Economics and Sociology* **46**(4): 459-472.

Suzuki, S. and E. Akiyama (2007). Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. *Journal of Theoretical Biology* **245**(3): 539-552.

Suzuki, S. and E. Akiyama (2007). Three-person game facilitates indirect reciprocity under image scoring. *Journal of Theoretical Biology* **249**(1): 93-100.

Swenson, W., D. S. Wilson and R. Elias (2000). Artificial ecosystem selection. *Proceedings of the National Academy of Sciences of the United States of America* **97**(16): 9110-9114.

Takahashi, T. (2007). Non-reciprocal altruism may be attributable to hyperbolicity in social discounting function. *Medical Hypotheses* **68**(1): 184-187.

Tesfatsion, L. (2002). Agent-Based Computational Economics: Growing Economies From the Bottom Up. *Artificial Life* **8**(1): 55-82.

Tesfatsion, L. (2003). Agent-based computational economics: modelling economies as complex adaptive systems. *Information Sciences* **149**(4): 262-268.

Skyes J.B. Eds. (1976).*The Concise Oxford dictionary of current English : based on the Oxford English dictionary and its supplements* 6th ed. Clarendon Press.

Tomer, J. F. (2007). What is behavioural economics? *Journal of Socio-Economics* **36**(3): 463-479.

Tooby, J. and L. Cosmides (1992). The Psychological Foundations of Culture. In *The Adapted Mind: Evolutionary psychology and the generation of culture*. J. Barkow, L. Cosmides and J. Tooby. Eds. Oxford University Press, New York**:** 19-136.

Tracer, D. P. (2003). Selfishness and Fairness in Economic and Evolutionary Perspective: An Experimental Economic Study in Papua New Guinea. *Current Anthropology* **44**(3): 432-438.

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology* **46**(1): 35-57.

Tucker, W. T. and S. Ferson (2008). Evolved Altruism, Strong Reciprocity, and Perception of Risk. *Annals of the New York Academy of Sciences* **1128**(Strategies for Risk Communication Evolution, Evidence, Experience): 111-120.

Vernon, L. S. (1976). Experimental Economics: Induced Value Theory. *The American Economic Review* **66**(2): 274-279.

Vine, I. (1983). Sociobiology and social psychology--rivalry or symbiosis? The explanation of altruism. *British Journal of Social Psychology* **22**(1): 1-11.

Von Neumann, J. and O. Morgenstern (1944) *Theory of Games and Economic Behavior*. Princeton University Press, New Jersey.

Voss, T. (2001). Game theoretical perspectives on the emergence of social norms. In *Social Norms*. M. Hechter and K. D. Opp. Eds. Russell Sage Foundation, New York**:** 105–138.

Walker, J. and E. Ostrom (2007). *Trust and Reciprocity as Foundations for Cooperation: Individuals, Institutions, and Context*. Indiana University, Bloomington**:** 1-43.

Wedekind, C. and M. Milinski (2000). Cooperation Through Image Scoring in Humans. *Science* **288**(5467): 850-852.

Werhane, P. H. (2000). Business Ethics and the Origins of Contemporary Capitalism: Economics and Ethics in the Work of Adam Smith and Herbert Spencer. *Journal of Business Ethics* **24**(3): 185-198.

West, S. A., A. S. Griffin and A. Gardner (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology* **20**(2): 415-432.

William, G.C. (1966). Natural Selection, the Cost of Reproduction, and a Refinement of Lack's Principle. *The American Naturalist* **100**(916):687-690.

Wilson, B. J. (2008). Language games of reciprocity. *Journal of Economic Behaviour & Organization* **68**(2): 365-377.

Wilson, D. R. (2006). The evolutionary neuroscience of human reciprocal sociality: A basic outline for economists. *Journal of Socio-Economics* **35**(4): 626-633.

Wilson, D.S., D.T. O'Brien and A. Sesma (2009). Human prosociality from an evolutionary perspective: variation and correlations at a city-wide scale. *Evolution and Human Behavior* **30**(2009): 190–200.

Wilson, D.S. and E.O. Wilson (2007). Rethinking the Theoretical Foundation of Sociobiology. *The Quarterly Review of Biology* **82**(4): 327-348.

Wilson, E. O. (2000). *Sociobiology: The New Synthesis*. Belknap Press, Cambridge.

Wischniewski, J., S. Windmann, G. Juckel, and M. Brüne (2009). Rules of social exchange: Game theory, individual differences and psychopathology. *Neuroscience & Biobehavioural Reviews* **33**(3): 305-313.

Wood, J. C. (1993). *Adam Smith: Critical Assessments*. Taylor & Francis, Oxford.

Wuthnow, R. (1993). Altruism and Sociological Theory. *The Social Service Review* **67**(3): 344-357.

Yang, C. -L., C. -S. J. Yue and I. -T. Yu (2007). The rise of cooperation in correlated matching prisoners dilemma: An experiment. *Experimental Economics* **10**(1): 3-20.

Yao, X. (1996). Evolutionary stability in the n-person iterated Prisoners' Dilemma. *Biosystems* **37**(3): 189-197.

Yu, R. and X. Zhou (2007). Neuroeconomics: Opening the black box behind the economic behaviour. *Chinese Science Bulletin* **52**(9): 1153-1161.

Zak, P. J. (2007). The Neuroeconomics of Trust. In *Renaissance in behavioural economics: essays in honor of Harvey Leibenstein*. R. S. Frantz and H. Leibenstein. Eds. Routledge, New York.

Zak, P. J., R. Kurzban and W. T. Matzner (2005). Oxytocin is associated with human trustworthiness. *Hormones and Behaviour* **48**(5): 522-527.

ATTACHMENT A: PROCESS STEPS FOR THE TWO-PERSON RANDOM INTERACTION MODEL (2PRIM)

Two-person random interaction model (2PRIM)

The two-person random interaction model (2PRIM) has been developed and simulated in *MATLAB* R2008B and it follows the following process:

1. At the beginning of the experiment a set of 1000 individuals are chosen at random (based on a uniform discrete distribution) from a pool of N individuals.
2. Each of these 1000 players is provided a random assignment along the A/S continuum (a continuous uniform distribution). Each individual thus has a randomly assigned Selfishness and Altruism trait value. The Altruistic trait value equals 1 – Selfish trait value. Therefore, the level of altruism plus the level of selfishness for each individual equals 1.
3. For each experiment, 1000 games that are run. For each game, there are 100,000 rounds undertaken.
4. At the beginning of each round, two individuals will be selected at random from this pool of 1000 individuals (Note: a single individual cannot play against themselves, so there will always be two different individuals).
5. Each individual would already have a Selfish and Altruistic trait value that was assigned to them at random based on the A/S continuum (as per point 2 above).
6. The utility for each player is then calculated according to equation 1.
7. Two new players will be selected at random from this pool of 1000 individuals for the next round.
8. Resultantly, as these two players are selected at random – some players could end up playing more rounds than others that replicates the concept of luck which we observe in human life.

9. The resource pool of 1000 players will then be sorted by utility values in ascending order and the bottom 10 per cent of players in this sorted list will be eliminated from the game in order to increase variety in the population.
10. An additional 10 per cent of players will be selected at random from the pool of N individuals, from where the initial 1000 individuals were selected.
11. The utility values of each player (in the pool of 1000 individuals) will be reset to zero at the beginning of each game.
12. This process will continue until the 1000 games are completed.

Two-person random interaction model (2PRIM) INCORPORATING PUNISHMENT

In chapter 6, this two-person random interaction model (2PRIM) was modified further to include punishment and the following process is followed when punishment is included:

1. At the beginning of the experiment a set of 1000 individuals are chosen at random (using a simple random process) from a resource pool.
2. Each of these 1000 players is provided a random assignment along the A/S and P/NP continuum. Each individual thus has a Selfishness and Punishment trait value, which the randomly assigned A/S and P/NP values respectively. The Altruistic trait value equals 1 – Selfish trait value.
3. For each experiment, there are 1000 games that are run. For each game, there are 1,000,000 rounds undertaken and each round has 2 phases.
4. At the beginning of each round, two individuals will be selected at random from this pool of 1000 individuals (Note: a single individual cannot play against themselves, so there will always be two different individuals).
5. Each individual would already have a Selfish and Altruistic trait value that was assigned to them at random based on the A/S continuum (as per point 2 above).

6. These two individuals will play a one-off dual phase round before their utility is updated and they are placed back into the resource pool.
7. In phase one, the utility (payoff) of each player will be calculated based on their Altruism/Selfishness (A/S) trait value using the equation below:

$$\frac{A_1 + A_2}{2} \cdot CGfactor + S_1 \cdot SFfactor + (S_1 + S_2) \cdot Cfactor$$

Where: A_1: altruistic investment by Player 1 based on the A/S continuum.

A_2: altruistic investment by Player 2 based on the A/S continuum

S_1: selfish investment by Player 1 based on the A/S continuum

S_2: selfish investment by Player 2 based on the A/S continuum

CGfactor = explains the amount the Common Pool amount will be multiplied by at the end of each round in the game. Standard value of the CGfactor = 1, based on constant returns to scale.

SFfactor = represents the degree of utility that a selfishness person receives when taking a Selfish action. An increase in level of guilt decreases the level of utility received which is represented by a decrease in SFfactor. Standard value of the SFfactor = 1, based on constant returns to scale.

Cfactor = the cost that players pay for destructive competition. When, players are selfish they reduce the overall value of the Prisoners' Dilemma game pay-off. Standard value of the Cfactor = 0.25, based on constant returns to scale.

8. In the second phase of this round, each player will have a P/NP value (that was assigned to them in the resource pool at random in point 2 above – similar to how the A/S values were assigned). The P/NP value shows the tendency of that player to punish their opponent.

9. Only the less Selfish individual can punish the more Selfish individual. Further, the P/NP value of the punisher needs to be greater than a probabilistic amount (ascertained using a random number generator), only then can that punisher punish his opponent.
10. Punishment happens by the punisher increasing their selfishness by the Pfactor in the second phase of the round.
11. However, this punisher also needs to pay a cost to punish (i.e. Pcost) because he is increasing his selfishness in the second round to a level that is unnatural to him (therefore, this psychological cost will equal Pcost).
12. Based on the equation below each player's utility will be calculated at the end of the second phase. The utility function now becomes:

$$(\frac{A_1 + A_2}{2} \cdot CGfactor) + (S_{X1} \cdot SFfactor) + ((S_{X1} + S_2) \cdot Cfactor) - (PCost) \quad (2)$$

Where,

$$S_{X1} = S_1 + Pfactor \text{ (where } 0 < S_{X1} < 1) \quad (3)$$

Punishment is applied probabilistically in 2PRIM – where a random variable is used to assess if punishment will be applied. If this random variable (i.e. generated using a random number generator) is greater than the Pfactor value then punishment will be applied that will equal the value of Pfactor, else punishment will not be applied.

Where: *Pfactor:* the level by which the selfishness of the less selfish player will increase, if the punishment factor of this less selfish player is greater than a random value. Standard value of the Pfactor = 0.25, based on constant returns to scale.

Pcost: the cost faced by the punisher as they increasing their level of selfishness (by adding Pfactor) to a level that is unnatural to them. Standard value of the Pcost = 0.10, based on constant returns to scale.

13. The final utility for each player at the end of the two-phase round will be equal to their utility at the end of the second phase.
14. At the end of this two-phase round, these two players will be returned to the pool of the 1000 players and two new players will be selected at random from this pool. Resultantly, as these two players are selected at random – some players could end up playing more rounds than others, which includes the concept of luck that occurs with human beings in everyday life.
15. After one game (100,000 rounds) has been played, players will have different utility values as they have different A/S and P/NP values plus they would have probably played different number of rounds.
16. The resource pool of 1000 players will then be sorted by utility values in ascending order.
17. Then, the bottom 10 per cent of the players in this sorted list (by utility values) will be eliminated from the game in order to increase variety in the population.
18. A same number (10 per cent) of new players will be accepted in the resource pool at random and their A/S and P/NP values will be set at random as well.
19. The utility values of each player (in the pool of 1000 individuals) will be reset to zero at the beginning of each game.
20. This process will continue until the 1000 games are completed.

ATTACHMENT B: *MATLAB* CODE OF MODEL (2PRIM) DEVELOPED IN CHAPTERS 5 AND 6

This is the model for chapter six (2PRIM with punishment) and it builds on the code used for the model in chapters five (2PRIM), which has not been provided (the code for chapter five is wholly incorporated in the code used for chapter six below). The code used for the simulation of the model in chapter five is available on request.

```
N = 1000;                       %number of players
T = 100000;                     %number of rounds in each game
NoG = 10;                       %number of games
maxSF_CG = 2;
SFIncrement = 0.1;
CGIncrement = 0.1;
SFfactor = -SFIncrement;        %selfish factor - also the guilt factor
Cfactor = 0.25;                 %competition factor
Pfactor = 0.50;                 %punishment factor
Pcost = 0.1;                    %cost of punishment
UtilityVector = [];
SelfishVector = [];
UtilityMatrix = [];
SelfishMatrix = [];
```

```
for d = 0:SFIncrement:maxSF_CG
    SFfactor = SFfactor + SFIncrement;
    UtilityVector = [];
    SelfishVector = [];
    CGfactor = -CGIncrement;
    for c = 0:CGIncrement:maxSF_CG                          %TT numbers of games will be played
        CGfactor = CGfactor + CGIncrement;
        X = [rand(N,1) rand(N,1) zeros(N,1)];               %create matrix X with N rows & 3 columns
        for b = 0:NoG
            X(:,3) = zeros(N,1);
            for a = 1:T
                PL1 = floor(1 + (N - 1)*rand(1));           %player1
                PL2 = floor(1 + (N - 1)*rand(1));           %player2
                if PL1 == PL2
                    PL2 = floor(1 + (N - 1)*rand(1));       % re pick player 2 to ensure it is different from player 1
                end

                %calculate utility at the end of round 1
                utility1 = ((1 - X(PL1,1)) + (1 - X(PL2,1)))*CGfactor/2 + (SFfactor*X(PL1,1))-((X(PL1,1)+X(PL2,1))*Cfactor);
                %Payoff for Player A after Round 1
                utility2 = ((1 - X(PL1,1)) + (1 - X(PL2,1)))*CGfactor/2 + (SFfactor*X(PL2,1))-((X(PL1,1)+X(PL2,1))*Cfactor);
                %Payoff for Player A after Round 1
```

```
        %calculate utility at the end of round 2
          if X(PL1,1) > X(PL2,1) && X(PL2,2) > rand(1)
            NewSelfishness = X(PL2,1)^(Pfactor^X(PL2,1));
            utility1 = ((1 - X(PL1,1)) + (1 - NewSelfishness))*CGfactor/2 + (SFfactor*X(PL1,1))-((X(PL1,1)+NewSelfishness)*Cfactor);
         %Payoff for Player A after Round 2
            utility2 = ((1 - X(PL1,1)) + (1 - NewSelfishness))*CGfactor/2 + (SFfactor*NewSelfishness)-((X(PL1,1)+NewSelfishness)*Cfactor)
- Pcost;                    %Payoff for Player A after Round 2
         elseif X(PL2,1) > X(PL1,1) && X(PL1,2) > rand(1)
            NewSelfishness = X(PL1,1)^(Pfactor^X(PL1,1));
            utility1 = ((1 - NewSelfishness) + (1 - X(PL2,1)))*CGfactor/2 + (SFfactor*NewSelfishness)-((NewSelfishness+X(PL2,1))*Cfactor)
- Pcost;                    %Payoff for Player A after Round 2
            utility2 = ((1 - NewSelfishness) + (1 - X(PL2,1)))*CGfactor/2 + (SFfactor*X(PL2,1))-((NewSelfishness+X(PL2,1))*Cfactor);
                %Payoff for Player A after Round 2
         end

         %update utility in X after round 2
         X(PL1,3) = utility1 + X(PL1,3);                    %overwrite Utility for player 1
         X(PL2,3) = utility2 + X(PL2,3);                    %overwrite Utility for player 2
        end
```

```
            X = sortrows(X,3);                                  %sort rows based on utility
            X(1:N/10,:) = [];                                   %delete 10% of individuals
            Xadd = [rand(N/10,1) rand(N/10,1) zeros(N/10,1)];   %add new 10% of individuals at random
            X = [X; Xadd]; %#ok<AGROW>                          %add new 10% individuals at random to the end of matrix X
        end
        UtilityVector = [UtilityVector mean(X(:,3))]; %#ok<AGROW>
        SelfishVector = [SelfishVector mean(X(:,1))]; %#ok<AGROW>
    end
    UtilityMatrix = [UtilityMatrix; UtilityVector]; %#ok<AGROW>
    SelfishMatrix = [SelfishMatrix; SelfishVector]; %#ok<AGROW>
end
X(N-N/10:N,:) = [];                                             %delete newly added rows with zero utility
SFVector = 0:SFIncrement:2;
CGVector = 0:CGIncrement:2;

figure (1)
mesh (SFVector, CGVector, UtilityMatrix);
xlabel('Selfishness Factor')                                    %set x-axis label
ylabel('Common Good Factor')                                    %set y-axis label
zlabel('Level of Utility')                                      %set z-axis label
title('Change Utility with varying Selfish & Common Good Factors')   %set chart title
```

```
figure (2)
mesh (SFVector, CGVector, SelfishMatrix);
xlabel('Selfishness Factor')                                          %set x-axis label
ylabel('Common Good Factor')                                          %set y-axis label
zlabel('Level of Selfishness')                                        %set z-axis label
title('Change Selfishness with varying Selfish & Common Good Factors')   %set chart title
```