Individuals' Behaviour in Social Dilemma Games and the Role Played by Persuasion. Theory and Experiments.

Maria Vittoria Levati

DPhil in Economics

University of York

Department of Economics and Related Studies

April 2000

Supervisor: Professor J. D. Hey

Abstract

This work is a study of individuals' decisions pertaining to social dilemma games. Its main purposes are to propose a new explanation—the *persuasion hypothesis*—for the cooperative behaviour which people (beyond the game theoretic prediction) are found to exhibit in these kinds of games, and to investigate its empirical strength by means of a series of experiments. These experiments are specifically designed to assess the validity of my hypothesis in comparison with alternative models of cooperative behaviour.

Two main (contrasting) observations can be drawn from such an empirical investigation: persuasion appears to be a plausible explanation for previously inexplicable cooperation in a simple linear setting where the subjects' decisions are binary, but it fails to explain the data from a more complicated non–linear setting where the solutions lie in the interior of the strategy space.

Contents

Accompanying material ix

Acknowledgements xi

1 Introduction 1

2 Game theory and social dilemma games 6

2.1 Introduction 6

2.2 Game theoretic approach to individuals' interactions 7

2.2.1 The assumptions of game theory 8

2.2.2 The elements of game theory 11

2.3 Social dilemma games: definition and relevance 17

2.3.1 Introduction to the logic of social dilemmas 17

2.3.2 The public goods (free–riding) problem 18

2.3.3 Game theoretic predictions 19

2.3.4 How can the public goods problem be theoretically solved? 21

3 How people actually behave in dilemma game situations: a first generation of experiments 25

3.1 Introduction 25

3.2 Experimental methods in economics 26

3.2.1 Why laboratory experiments? Their main purposes and advantages 28

3.2.2 Realism and controlled economic experiments 30

3.3 First attempts to estimate demand for public goods 34

3.4 Psychological approach to social dilemmas 36

3.5 First systematic approaches to the public goods problem 39

3.6 Reaction of the economists: some first responses 42

3.7 An important final question: Is it really the theory that is wrong? 44

4 Why do people cooperate if full rationality demands defection? Some hypotheses of cooperative behaviour **48**
4.1 Introduction 48
4.2 A taxonomy of some hypotheses of cooperative behaviour 48
4.2.1 Dividing the dominant strategy argument into three parts 49
4.2.2 Altruism and equity theories: alternatives to motive . . . 49
4.2.3 Reciprocity and commitment theories: alternatives to choice 54
4.2.4 Reputation building, errors and learning theories: alternatives to cognition 56
4.3 How well do these theories perform in laboratory settings? 60
4.3.1 Repetition, learning and strategies 61
4.3.2 The noise versus altruism hypotheses 67
4.3.3 Framing, altruism, reciprocation and distributive concerns 79
4.4 Summary and conclusions 88

5 The persuasion hypothesis: what it is and how it works **90**
5.1 Introduction 90
5.2 The persuasive player: an automaton with a bit of rationality . . 91
5.2.1 The persuasive player's rational calculus 92
5.2.2 Minimum periods of cooperation necessary for identifying a persuasive player 97
5.2.3 The relationship between a persuader's contribution and the others' behaviour 98
5.3 Game–theoretic consequences of the persuasive strategy 100
5.4 The persuaded player 103
5.5 A comparison with other theories of cooperative behaviour . . . 105
5.5.1 Theories of commitment 105
5.5.2 Theories of reciprocity 107
5.5.3 Theories of altruism 109
5.5.4 Reputation building/strategies hypothesis 111
5.6 Final thoughts 114

6 My 'simple' experiment **115**
6.1 Introduction 115
6.2 The constituent game and the decision round's structure 116
6.3 Experimental parameters and procedures 120
6.3.1 About the first subsession 120
6.3.2 About the second subsession 121
6.3.3 Subject pool 122

6.4 Constructing the alternative hypotheses 122
6.4.1 The persuasive type 123
6.4.2 The reciprocal type 125
6.4.3 The strategic type 125
6.4.4 The Kantian type 126
6.4.5 The altruistic type 126
6.4.6 The persuaded type 126
6.4.7 The Nash type 127
6.4.8 Summary of the criteria used in identifying a player type 127
6.5 Aggregate results 128
6.6 Individual results 131
6.6.1 Testing subjects' confusion 131
6.6.2 Testing the alternative hypotheses of cooperative behaviour 133
6.7 Concluding discussion 144

7 A deeper experimental analysis: towards more decisive conclusions 146
7.1 Introduction 146
7.2 The voluntary contribution mechanism environments 147
7.2.1 The *basic* non linear game 147
7.2.2 The *separation* game 148
7.3 Experimental design: parameters and procedures 150
7.3.1 The endogenous questionnaire and the non questionnaire treatment 151
7.3.2 The communication treatment 152
7.3.3 Subject pool 153
7.4 Persuasion and other theories of cooperative behaviour 154
7.4.1 Persuasive behaviour theory 154
7.4.2 Commitment theories 156
7.4.3 Reciprocity theories 157
7.4.4 Altruism theories 158
7.5 Aggregate results 159
7.5.1 Treatment effects on aggregate behaviour 159
7.5.2 Interior versus boundary equilibria 166
7.6 Individual results 169
7.6.1 The phase effect at individual level 170
7.6.2 The cutoff period θ 171
7.6.3 Comparing decisions with others' behaviour 172

7.6.4 Comparing an individual's expectations with the others' contributions . . . 184
7.7 Conclusions . . . 185

8 Summary and suggestions for future research 187
8.1 Introduction . . . 187
8.2 Theoretical considerations . . . 187
8.3 Achievements to date . . . 189
8.3.1 On the strengths and weaknesses of my design . . . 190
8.3.2 Major results: a summary . . . 191
8.4 Other theories and other empirical studies . . . 194
8.5 Research agenda . . . 195
8.6 Conclusions . . . 196

A First experiment's instructions 198
A.1 Instructions for the first subsession . . . 198
A.2 Instructions for the second subsession . . . 203

B Second experiment's instructions 204

C First experiment's data 211
C.1 First subsession's data . . . 212
C.2 Second subsession's data . . . 220

List of Figures

6.1 Form in which the constituent game was presented to subjects. . 118
6.2 Average of subjects cooperating in each phase. 129

7.1 Kantian constrained utility in the separation game. 156
7.2 Altruist's reaction function in the separation game. 158
7.3 Average individual contributions displayed separately in the 3 phases of each treatment. 160
7.4 Average contributions in the three treatments for each phase. . . 161
7.5 Average contributions in the three phases for each treatment. . . 162

List of Tables

3.1 Dawes et al.'s (1977) experiment: Payoff matrix. 37
3.2 Dawes et al.'s (1977) experiment: Proportion of subjects defecting. 38
3.3 Marwell and Ames' (1979) experiment: Payoff from group exchange in large, unequal–interest, unequal–resource groups. . . . 40
3.4 Isaac et al.'s (1984) experiment: Percentage contribution data. . 44

4.1 Prisoner's dilemma payoff matrix. 53
4.2 Andreoni's (1988) experiment: Average investment in public good per subject. 62
4.3 Burlando and Hey's (1997) experiment: Differences in percentages dumped before and after the restart. 65
4.4 Andreoni's (1995a) experiment: Percentage of subjects contributing zero to the public good per round. 69
4.5 Classification of BBK's hypotheses by key concept and strategy sensitivity. 81
4.6 BBK's (1997) experiment: Predictions about player A's mean contributions. 83

6.1 z–statistics for the differences in cooperators' proportions between phases. 130
6.2 Distribution of subjects across the three phases of subsession 1. . 134
6.3 Classification of each subject according to his type across the three phases of subsession 1. 136
6.4 Distribution of subjects across the three phases of subsession 2. . 138
6.5 Classification of each subject according to his type across the three phases of subsession 2. 139
6.6 The 'restart effect' in this experiment. 140

7.1 Value of the reciprocity–test–period μ corresponding to each possible contribution level $\bar{g}_p$ in the basic game. 155

7.2 Differences in individual average contributions between the CT and the two treatments without communication. 163
7.3 Differences in individual average contributions between the NCT and the NQT. 165
7.4 z–statistics for the differences between phases under each treatment. 167
7.5 Percentages of Nash equilibrium strategies observed in each phase of each treatment. 167
7.6 Percentages of non–equilibrium strategies. 168
7.7 Number of cases in which the phase effect is significant. 170
7.8 Characteristics of each subject's pooled regression under the NCT. 174
7.9 Characteristics of the regression for the subject whose phase effect is significant under the NCT. 175
7.10 Characteristics of each subject's selected regression in the second phase of the NCT. 176
7.11 Characteristics of each subject's pooled regression under the CT. 177
7.12 Characteristics of each subject's regression when the phase effect is significant under the CT. 178
7.13 Characteristics of each subject's selected regression in the second phase of the CT. 179
7.14 Characteristics of each subject's pooled regression under the NQT. 180
7.15 Characteristics of each subject's regression when the phase effect is significant under the NQT. 181
7.16 Characteristics of each subject's selected regression in the second phase of the NQT. 182

A.1 First question of the questionnaire 202
A.2 Second question of the questionnaire 203

B.1 The payoff matrix in Phases 1 and 3. 209
B.2 The payoff matrix in Phase 2. 210

Accompanying material

One computer disc accompanies the thesis. It contains all individual data collected from the experiment described in Chapter 7.

I preferred to provide the data from Chapter 7's experiment in a disk rather than include them in an Appendix of the thesis (as I did for the data from Chapter 6's experiment) because of their quantity: 72 subjects took part in this experiment, and each subject made 6 decisions in each of 30 periods.

The data can be read with any text editor, and they are separated according to experimental session. Since the experiment consisted of three sessions—each of which employed one of three treatments—I have created three directories and named them CT (for communication treatment), NCT (for non–communication treatment), and NQT (for non–questionnaire treatment).

Each directory includes the data files of the 24 people who participated in the session–treatment to which the directory refers (so that a total of $24 \times 3 = 72$ individual data files are provided).

In each individual file, there are three tables which pertain to the three phases (i.e., supergames) into which each session–treatment was divided.

In the directories CT and NCT, each of these three tables has six columns and ten rows. The latter refer to the number of periods for which the game was repeated in each phase. The content of each column is the following:

i) the first column reports the contribution decisions of the subject;

ii) the second column reports the motivations he provided for the corresponding period–decision;

iii) the third column reports his guesses of the motivation underlying the second group member's period–decision;

iv) the fourth column reports his guesses of the motivation underlying the third group member's period–decision;[1]

v) the fifth column reports his predictions about the next contribution decision

[1] I strongly recommend to look at pages 206 and 207 in Appendix B while reading the data in columns 2–4, since the lists of nine alternatives among which subjects had to pick out one in order to answer the questions to which columns 2–4 refer are described on these pages.

of the second member of his group;
vi) finally, the sixth column reports his predictions about the next contribution decision of the third member of his group.

In the directory called NQT, each of the three tables showed in each individual file has only one column, which reports the contribution decisions of the subject.

For each individual file and for each of the three phases in it, I also indicate who—among the 24 participants—were the two partners of the subject.

Acknowledgements

This work would not have been possible without the help of several institutions and people. First, I should like to thank the Economic and Social Research Council (ESRC) and the European Network for the Development of Experimental Economics and its Application to Research on Institutions and Individual Decision Making (ENDEAR) for funding the experiments reported in Chapters 6 and 7 with the grant numbers R000236636 and FMRXCT98 0238, respectively.

I would like also to thank my former colleague at the EXEC Marie–Edith Bissey for stimulating discussions and for a very pleasant working climate.

I am also indebted to Anastasios Koukoumelis for carefully reading through the entire manuscript, pointing out some flaws, and providing me with a lot of helpful comments.

In particular, however, I owe a debt of gratitude to my supervisor Professor John Hey who continuously encouraged my research activity, much more than one could usually expect.

Finally, I have to thank my parents for their never–ending patience and their moral support while writing this thesis.

York, April 2000

Chapter 1

Introduction

My intention in this work is to study people's behaviour in social dilemma settings with the aim of identifying forces and underlying principles behind their actions. Social dilemmas are defined by two simple properties: *a*) each individual receives a higher payoff for a socially defecting choice than for a socially cooperative one, no matter what the other individuals in the society do; but *b*) all individuals are better off if all cooperate rather than if all defect.

The specific social dilemma to which the major part of this work is devoted is the public goods problem.

A public good has two essential attributes: non–rivalry in consumption and non–excludability. The former characteristic refers to the possibility of simultaneous consumption of the same unit of the good by multiple consumers. The other characteristic means that it is difficult to prevent consumption of the good by those who fail to pay. Hence, when the public good must be financed by private arrangements, any individual has an incentive to free–ride on the contributions of the others. If everyone follows such an individual incentive, the good is provided at an inefficient level; i.e., at a Nash equilibrium level which is Pareto dominated by the non–equilibrium outcome where everyone contributes to the public good.

The incentive to free–ride has generated a standard presumption among economists that decentralised allocation mechanisms cannot be relied upon to provide public goods efficiently (e.g., Samuelson (1954)). The vast resources devoted to the government provision of goods and services in developed economies support a general perception that the free–rider problem is pervasive. This presumption, however, is not without critics. The claim is that government provision of public goods, financed by lump–sum taxes, is not necessarily more efficient than private provision of public goods (e.g., Bergstrom et al. (1986)). The public broadcasting service (PBS) in the United States, for instance, is

a private, non–profit corporation whose members are public TV stations, and which is almost entirely financed by voluntarily contributions. In contrast, public TV stations in Canada are mainly financed by taxes. A comparison between the two stations shows that the latter does less well than the former both in quality and budget. Noting the large number of Canadian underwriters to United States' PBS, one has evidence that Canadians are actually willing to make voluntarily contributions to finance quality broadcasting.

To address the question of the significance of the free–rider hypothesis (i.e., that the provision of public goods will not occur at all or will be suboptimal) experimenters in economics as well as in psychology started investigating human behaviour in well controlled public goods situations in the laboratory.

What emerges from these early experiments is that—contrary to the theoretical predictions—people are willing to cooperate. Generally speaking, the major findings of this vast literature are the following: 1) in one shot trials and in the initial stages of finitely repeated trials, subjects contribute around halfway between the Pareto–efficient level and the free–riding level; 2) contributions decline with repetition; 3) face–to–face communication improves the rate of contributions.[1]

The research problem becomes, hence, to discover *when* subjects cooperate and *why* they do so.

The aim of this work is (a) to give an overview of the various theories that have been advanced for explaining individuals' cooperative behaviour; (b) to present a new hypothesis for such a behaviour; (c) to investigate how well the former theories perform in laboratory settings and how much of the previous inexplicable cooperation can be explained by my new hypothesis.

As for point (a), since the full rationality argument for playing the dominant strategy can be divided into three components: *motive* (a player prefers to have more money to having less), *cognition* (a player can identify the action that best satisfies motive), and *choice* (a player chooses the action that best satisfies motive), and since most theories dealing with dilemma games depart from full rationality on one of the three components and rely on it (sometimes implicitly) for the other two, in this work, I will classify the existing theories of cooperative behaviour by the major component of deviation. In particular, altruism and equity theories will be regarded as alternative to motive; reciprocity and commitment theories as alternatives to choice; reputation building/strategies

[1]Extensive surveys of the early experimental studies on voluntary contributions to public goods are presented by Ledyard (1995) and Davis and Holt (1993).

hypothesis, confusion and simple learning as alternatives to cognition.

Examination of several experiments which have been run in order to verify the descriptive power of these theories reveals that the issue of which approach best explains the data remains open. Each of the models considered have been shown to effectively explain outcomes in only some subset of the various experimental settings. In other words, no model appears to be able to account for all observed cooperation. This suggests that some of the reasons for people voluntarily contributing to public goods remain to be found. Accordingly, I propose a new behavioural hypothesis through which I try to explain previously inexplicable cooperation.

This hypothesis relies on *persuasion* as a valid principle organising behaviour, where persuasion is defined as the intention to influence the others in order to make them behave in a way that they otherwise would not. In settings like voluntary contribution mechanism environments, the persuasion hypothesis rests on two facts: 1) an agent realises that his own payoff is higher when other group members contribute to the public good an amount greater than the equilibrium rather than the equilibrium amount; and consequently 2) in order to get these higher benefits, (if the mechanism is implemented in a sequence of periods) he tries to push the others (who are free–riding) towards contribution decisions.

When no kind of verbal communication is allowed, the only way that a player has got to signal information to the others is by his choices of moves as the game proceeds. Thus, he chooses to contribute a constant amount for some consecutive periods (even if the others are making the socially defecting choice) because he hopes that through this choice he can *either* educate the free–riders about the mutual benefits connected to contribution, *or* reduce their uncertainty about the way in which he (himself) will play the game (by raising their expectations about his own contributions). In both cases, it is the hope of achieving the most favourable outcome for the group (and, hence, *for himself*) that, under the persuasion hypothesis, leads an agent to select a cooperative arrangement. In this sense, although the persuasion hypothesis is in the spirit of Kant's categorical imperative,[2] it differs from Kant's moral because, under my hypothesis, contributions are not conceived as obligations independent of any strategic considerations. A persuasive player must, in fact, believe that, during the interaction process, his group members are willing to modify their

[2] A persuasive agent is, in fact, expected to act in accordance with collective rationality, which prescribes a course of action to all players simultaneously, and contrary to individual rationality, which prescribes to each player the course of action most advantageous to him under the circumstances.

initial views (i.e., they are available to be persuaded). For this reason, after some periods of unconditional cooperation, he looks at what the others do and keeps on the same contribution decision only if this is reciprocated.

A similar reasoning implies that, if he is successful in his attempts to push the partners towards his own's contribution, the persuasive player perseveres with this choice as long as he observes that each of the others does the same; otherwise, he modifies it in the direction of the others' average contribution in the previous period. There exists, hence, the possibility for him to free–ride if other players in his group are free–riding. The persuasive behaviour theory appears, therefore, to be compatible with both the observed successes and the observed failures of voluntary cooperation.

Since my interpretation of individuals' cooperative behaviour generates testable predictions which differ from those of earlier theories, its capability of accounting for previously inexplicable cooperation can be empirically verified. Thus, I shall conduct a series of public goods experiments whose designs allow for a direct test of the persuasive behaviour hypothesis and for its comparison with earlier theories of cooperative behaviour.

The experiments combine behaviour with belief–elicitation and (what is less common) with an investigation into subjects' motivations for their own decisions. Asking subjects to motivate their decisions is not only a very straightforward way to obtain insights into their decision rules, but it should also induce them to be more concentrated on the game, and to think seriously about the problem that they face.

Some previous experiments have attempted to investigate the relationship between an individual's decisions and his expectations about his fellow members' next decisions in public good settings. However, most have deceived subjects about the true contributions of the other players. In contrast to most of this previous literature, I will not use any deception. Indeed, I will elicit players' expectations by explicitly asking them to predict what the others will do in the upcoming period.

In a further departure from earlier experiments, I will elicit expectations in period t for period $t+1$ *after* a subject has observed the actual choices made by the others in t. On the one hand, this procedure avoids (or, at least, reduces) the psychological regularity known as 'falsus consensus effect', according to which individuals think others are like themselves; indeed, if a subject is asked to predict the decisions of the other members of his group once a decision has been already taken by them (revealing hence their attitude), he should be less inclined to believe that others' choices are like his. On the other hand, this

procedure allows subjects to answer all questions in a unique stage so as to divide each period into two parts only.

Outline of the thesis

The contents of the next seven chapters will be as follows.

- In Chapter 2, after presenting some relevant concepts of game theory, I shall introduce the specific types of interactive situations to which this work is devoted. I shall describe, hence, a typical public good game by emphasising the traditional game–theoretic predictions in it.
- In Chapter 3, I shall review the initial laboratory examination of the public goods problem. Before starting such a survey, I shall discuss some methodological issues concerning experimentation in economics.
- In Chapter 4, I shall concentrate on the explanations and justifications which the modern experimental research provides for the cooperative behaviour observed in the pioneering works.
- In Chapter 5, I shall present my 'persuasion' theory and I shall elaborate how persuasive behaviour can be used in public goods settings as instrument to achieve cooperation.
- The empirical analysis of the validity of my hypothesis and of its relevance in comparison with former theories of cooperative behaviour will start in Chapter 6. Here, as domain for this analysis, I shall use a simple three–player dilemma game with a standard linear public good specification whose theoretical solutions (the Nash equilibrium on one side, and the social optimum on the other) should be immediately clear to all subjects.
- In Chapter 7, the voluntary contribution mechanism environment in which people interact will be complicated by taking into account games with interior solutions which (according to various authors) introduce scope for potentially confounding effects.
- Chapter 8 concludes and discusses the potential contribution of this study to a positive theory of public good games.

Chapter 2

Game theory and social dilemma games

2.1 Introduction

The standard theory of choice, as developed by economists, is *rational–choice theory.* A main distinction that applies to rational choice situations is that between parametric and strategic decisions.

In a *parametric setting*, an agent faces external constraints that are in some sense given or parametric; first he estimates them as well as he can, and then he decides what to do. In such a setting an individual's choice is defined 'rational' if it is the choice which best satisfies his objectives.

In a *strategic setting*, individuals' decisions are interdependent. Before making up his mind, each agent has to anticipate what others are likely to do, which may require an estimate of what they anticipate that he will do; his decision enters as part–determinant of the constraints that shape his decision.

The branch of social science that studies strategic decisions and individuals' interaction is game theory. One can identify three types of strategic decisions and game theory: ideal–normative theory, prescriptive theory, and descriptive theory.

In *ideal–normative game theory* one assumes fully rational decision makers and also common knowledge of full rationality, and then analyses strategic behaviour under these conditions. "The assumptions are not realistic, but nevertheless ideal normative game theory is an important intellectual pursuit. The consequences of ideal normative game theory are of great philosophical significance" (Selten 1994, p. 1).

In *prescriptive game theory* the point of interest is what should be done if the participants in the game are not fully rational.

In *descriptive game theory* one is not interested in the topic how players should act, but how they actually act.

This chapter will be concerned only with ideal normative game theory. First, some relevant concepts of this approach and its cognitive foundations will be sketched. Then, the specific type of interactive situations to which this thesis is devoted (social dilemmas in general, and public goods games in particular) will be introduced. Finally, the normative game–theoretic predictions in these kinds of situations will be discussed.

The analysis of how people actually behave in social dilemma games and public goods experiments will be the topic of the next chapter.

2.2 Game theoretic approach to individuals' interactions

Game theory can be formally defined as "the study of mathematical models of conflict and cooperation between intelligent rational decision–makers" (Myerson 1997, p. 1).

In the past two decades, it has been increasingly adopted by several branches of the social sciences as well as by practical decision–makers. Two eminent game theorists explain this how follows: "Game Theory may be viewed as a sort of umbrella or 'unified field' theory for the rational side of social science... [it] does not use different, ad hoc constructs... it develops methodologies that apply in principle to all interactive situations" (Aumann and Hart 1992, p. 3).

A similar view is held by Elster (a well–known social theorist) who affirms: "If one accepts that interaction is the essence of social life, then... game theory provides solid microfoundations for any study of social structure and social change" (Elster 1982, p. 477).

All this interest is not difficult to understand. Modern game theory may be said to begin with the work of Zermelo (1913), Borel (1921), von Neumann (1928), and the great seminal book of von Neumann and Morgenstern (1944). In the language of game theory, a *game* refers to any social situation involving two or more individuals who have some understanding of how the outcome for one is affected not only by his own actions by also by the actions of all others. This is quite extraordinary. "From crossing the road in traffic, to decisions to disarm, raise prices, give to charity, join a union, produce a commodity, have children, and so on, it seems that we will be able to draw on a single model of analysis: the theory of games" (Hargreaves–Heap and Varoufakis 1995, p. 2).

As far as its specific solution techniques are concerned, the major successes

of game theory have come primarily from formalising common–sense intuitions in ways that allow analysts to see how such intuitions can be applied to fresh contexts and extended to slightly more complex formulations of situations. Game theory provided the few tools needed to frame and analyse theoretically unmanageable strategic behaviours in a mathematical theory.[1] This is predominantly how game theory makes a contribution: it helps to clarify some fundamental issues and debates in social life.[2] In this sense, its contribution is largely pedagogical. By repeating Selten's (1994) statement: "... ideal normative game theory is an important intellectual pursuit. The consequences of ideal normative game theory are of great philosophical significance".

The beginnings of this theory are presented here. Some formalism is necessary in such a presentation because the English language is not equipped enough to express the appropriate ideas compactly. Without some formalism it is, therefore, easy to get confused.

2.2.1 The assumptions of game theory

Rationality

An individual is *rational* if he makes decisions consistently in pursuit of his own objectives. In game theory (building on the fundamental results of decision theory) it is assumed that each player's objective is to maximise his expected utility.

This idea goes back to Bernoulli (1738), but its modern justification is due to von Neumann and Morgenstern (1947). Using remarkably weak assumptions about how a rational decision–maker should behave,[3] they show that for any rational decision–maker there exists a way of assigning utility numbers to the various possible outcomes that he cares about, such that he would always choose the option that maximises his expected utility. Such a result is called *expected–utility maximisation theorem.*

When there is uncertainty, expected utilities can be defined and computed only if all relevant uncertain events can be assigned probabilities, which quantitatively measure the likelihood of each event. Ramsey (1926) and Savage (1954) showed that, even where objective probabilities cannot be assigned to some events, a rational decision–maker should be able to act *as if* he knew all

[1] *The Economist* newspaper (24 December 1988) wrote about Jean Tirole: "Drawing on game theory and other strange techniques, [Tirole's] approach began to make sense of strategic behaviour that had seemed theoretically unmanageable".

[2] Cf., Kreps (1990); Myerson (1997); and Hargreaves–Heap and Varoufakis (1995).

[3] Particularly, the assumptions on the agent's preferences are the following: reflexivity, completeness, transitivity, continuity, non–satiation, and convexity.

the *subjective* probability numbers that are needed to compute these expected values.[4]

However, when we move from parametric to strategic settings (in which two or more decision–makers interact) a special difficulty arises in the assessment of subjective probabilities. For example, suppose that one of the factors that is unknown to some individual 1 is the action to be chosen by some other individual 2. To assess the probability of each of individual 2's possible choices, individual 1 needs to understand 2's behaviour, so 1 may try to imagine himself in 2's position. In this thinking process, 1 may realise that 2 is trying to rationally solve a decision problem on his own and that, in order to do so, he must assess the probabilities of each of individual 1's possible choices. Indeed, 1 may realise that 2 is trying to imagine himself in 1's position, to figure out what 1 would do. So the rational solution to each individual's problem depends on the solution to the other individual's problem. Neither problem can be solved without understanding the solution to the other.

The assumption of *common knowledge* is direct exactly to place some constraints on people's subjective expectations regarding the others' actions.

Common knowledge

The formulation of common knowledge was first given by Lewis (1969) in a philosophical study of conventions.

Aumann (1976) came up with the idea independently in a different context. By following him, we say that a fact is *common knowledge* among the players if every player knows it, every player knows that every player knows it, and so on; hence, every statement of the form "(every player knows that)^k every player knows it" is true for $k = 0, 1, 2, ...\infty$.

A player's *private information* is any information that he has that is not common knowledge among all the players in the game.

In general, whatever game model one may choose to study, the methods of game theory compel him to assume that this model must be common knowledge among the players. A game whose structure is common knowledge is called game with *complete information*.[5]

On the other hand, one can define a game with *incomplete information* as a game in which—at the first point in time when the players can begin to plan

[4]See also Friedman and Savage (1948).

[5]The structure of a game consists of its players, the decisions they face and the information they have when making them, how their decisions determine the outcome, and their payoffs for each outcome. The structure incorporates any repetition, correlating devices, or communication possibilities. (Cf., Gomes, Crawford and Broseta (1996))

their moves—some players have already private information about the game that others do not know. This initial private information is called the *type* of the player.

Games with incomplete information are studied by reducing them to games of complete information using a methodology introduced by Harsanyi (1967–68). For example, I may be in doubt whether my opponent at chess is really aiming to win or whether he wants to lose. Harsanyi would say that this should then not be seen as a two–player game with *perfect information* (where everybody knows what has happened previously in the game), but as a game with *imperfect information* with at least three players and where the opening event is a *chance move* that selects the type of my opponent.

A standard procedure is to take for granted that every player attaches the *same* probabilities to the possible outcomes of such a chance move.

It was Aumann (1976) who suggested that rational players will come to hold the same information. In particular, he asked whether rational players can *agree to disagree* by maintaining different probabilities for the same event, when (as happens in a game of incomplete information) the probabilities that they attach to the possible outcomes of chance events are common knowledge. His answer is that this is impossible. Intuitively, each player will use his knowledge of other players' estimates of the probabilities to refine his own estimate and this will continue until all the estimates are the same.[6]

However, Aumann requires a strong assumption in order to reach his result. He refers to it as the "Harsanyi doctrine". The idea is that before receiving any data, rational agents are all in the same position and therefore assign the same probabilities $prob(\omega) > 0$ to the states $\omega \in \Omega$. Moreover, these prior probabilities are common knowledge among the players. Hence, the Harsanyi doctrine is that *there is common knowledge of a common prior.* Later, the players have different experience which lead them to revise their probability. The current probabilities that they attach to the states $\omega \in \Omega$ are posterior probabilities obtained by Bayesian updating from the given common prior. And it is in this sense that it is common knowledge that these posterior probabilities are equal.

[6]See Bacharach (1985) for a more general expression of the idea.

2.2.2 The elements of game theory

The representation of games and some notation

The analysis of any game must begin with the specification of a model that describes the game. Several different forms are used, the most important of which are the *extensive* (or *dynamic* or *tree–diagram*) form and the *strategic* (or *normal*) form.

A *tree–diagram* shows the possible events that could occur in a game. The tree consists in a set of *branches* (or line segments), each of which connects two points that are called *nodes*. The first node in the tree is the *root* and represents the beginning of the game. The nodes that are not followed by any further branches are called *terminal nodes* and represent the possible ways in which the game may end. Each possible sequence of events that could occur in the game is represented by a path of branches from the root to one of these terminal nodes. When the game is actually played, the path depicting the actual sequence of events that will occur is called the *path of play*. The goal of game–theoretic analysis is to try to predict the path of play.

A simpler way to represent a game is to use the *strategic form*. To define a game in strategic form, we need only to specify the set of players in the game, the set of options available to each player, and the way that players' payoffs depend on the options that they choose.

Formally, a strategic form game is any Γ of the form

$$\Gamma = (N, (C_i)_{i \in N}, (u_i)_{i \in N}),$$

where N is the set of players; C_i is the set of (pure) strategies available to i $(\forall i \in N)$; and u_i is a function from $C = \times_{j \in N} C_j$ into the set of real numbers $\Re$. Here C denotes the set of all possible combinations or profiles of strategies that may be chosen by the various players, when each player j chooses one of his strategies in C_j. Given any strategy profile $c = (c_j)_{j \in N}$ in C, the number $u_i(c)$ describes the expected utility payoff that player i would get in the game if c were the combination of strategies implemented by the players. When a strategic form game is studied, it is usually assumed that the players all choose their strategies simultaneously; hence, there is no element of time in this analysis.

Domination and rationability

How should (or would) a rational person play a game? The simplest way proposed by game theorists suggests him to weed out strategies which are strate-

gically inferior by means of a step–by–step logic. Such elimination of strategies relies on what is called *dominance reasoning.*

To formally define dominance, we must first develop additional notation and illustrate the concept of best–response strategy.

For any player i, let C_{-i} denote the set of all possible combinations of strategies for the players other than i. Given any e_{-i} in C_{-i} and any d_i in C_i, (e_{-i}, d_i) is the strategy profile in C such that the i–component is d_i and all other components are as in e_{-i}.

Let $\Delta(C)$ be the set of all probability distributions over the set of strategy profiles for all players in Γ. If player i believes that some distribution η in $\Delta(C_{-i})$ predicts the behaviour of the other players, so each strategy combination e_{-i} has probability $\eta(e_{-i})$ of being chosen, then the rational player i wants to choose his own strategy in C_i to maximise his own expected payoff.

Player i's set of *best responses* to η is the set of i's strategies that maximise his payoff function. Formally, it is:

$$\underset{d_i \in C_i}{\operatorname{argmax}} \sum_{e_{-i} \in C_{-i}} \eta(e_{-i}) u_i(e_{-i}, d_i).$$

A *randomised strategy* for any player i is a probability distribution over C_i. Thus, $\Delta(C_i)$ denotes the set of all possible randomised strategies for i. For any pure strategy c_i in C_i and any randomised strategy σ_i in $\Delta(C_i)$, $\sigma(c_i)$ is the probability that i would choose c_i if he were implementing the randomised strategy $\sigma_i \in \Delta(C_i)$.

A strategy d_i in C_i is *strongly dominated* for player i iff there exists some randomised strategy σ_i in $\Delta(C_i)$ such that:

$$u_i(c_{-i}, d_i) < \sum_{e_i \in C_i} \sigma_i(e_i) u_i(c_{-i}, e_i), \quad \forall c_{-i} \in C_{-i}.$$

Alternatively, we can say that d_i is strongly dominated if and only if it can never be a best response for i, no matter what he may believe about the other players' strategies. This fact suggests that eliminating a strongly dominated strategy for any player i should not affect the analysis of the game, because player i would never use this strategy, and this fact should be evident to all other players if the hypothesis of common knowledge of rationality holds.

After one or more strongly dominated strategies have been eliminated from a game, other strategies that were not strongly dominated in the original game may become strongly dominated in the game that remains, and therefore they can be eliminated as well.

The process of successive elimination of strongly dominated strategies continues until no strategies can be further eliminated. Let $C_i^{(\infty)}$ denote the set of player i's strategies that remain after iterative elimination of strongly dominated strategies.

If D_i is a nonempty subset of C_i representing the set of strategies that player i might actually choose, then it can be shown that

$$D_i \subseteq C_i^{(\infty)}.$$

This means that, when all players are rational and there is common knowledge of rationality, no player can be expected to use any strategy that is iteratively eliminated by strong domination.

Thus, this first and weakest solution concept predicts only that the outcome of the game should be some profile of iteratively undominated strategies in $\times_{i \in N} C_i^{(\infty)}$. In many games, however, strong–dominance reasoning does not offer clear (or useful) predictions of what might happen.

One way to eliminate more strategies is to do iterative elimination of weakly dominated strategies.[7] But it is hard to argue that such a process would not affect the analysis of the game, because weakly dominated strategies could be best responses for a player if he feels confident that some strategies of other players have probability 0. Furthermore, the order in which weakly dominated strategies are eliminated may matter.[8]

Let us now suppose that all players choose their strategy independently. If the solution concept just introduced predicts that each player j will choose his strategy in the set D_j and if the assumption of common knowledge of rationality is true, then player i should choose a strategy that is best response to the other players' independent randomisation over their D_j sets. Let $H_i(D_{-i})$ be the set of such best responses. If all players are rational and this is common knowledge,

[7] A strategy d_i in C_i is *weakly dominated* for player i iff there exists some randomised strategy σ_i in $\Delta(C_i)$ such that:

$$u_i(c_{-i}, d_i) \leq \sum_{e_i \in C_i} \sigma_i(e_i) u_i(c_{-i}, e_i), \quad \forall c_{-i} \in C_{-i},$$

and, for at least one strategy combination $\hat{c}_{-i}$ in C_{-i}:

$$u_i(\hat{c}_{-i}, d_i) < \sum_{e_i \in C_i} \sigma_i(e_i) u_i(\hat{c}_{-i}, e_i).$$

[8] To avoid this order problem, Samuelson (1989) suggested that we might look for the largest sets $(D_i)_{i \in N}$ such that, for each player i, D_i is the set of all strategies in C_i that would not be weakly dominated for player i if he knew that the other players would only use strategy combinations in $D_{-i} = \times_{j \in N-i} D_j$. However, Samuelson himself showed that there exist games for which no such sets exist.

then the solution concept should satisfy:

$$D_i \subseteq H_i(D_{-i}), \quad \forall i \in N. \tag{2.1}$$

Bernheim (1984) and Pearce (1984) have shown that there exist sets $(D_j^*)_{j\in N}$ that satisfy (2.1) such that, for any other $(D_j)_{j\in N}$ that satisfies (2.1),

$$D_i \subseteq D_i^* \quad \forall i \in N.$$

The strategies in D_i^* are called *rationalisable strategies.* Any rationalisable strategy is a best response to some combination of independent randomisations over rationalisable strategies by the other players; that is,

$$D_i^* = H_i(D^*_{-i}), \quad \forall i \in N.$$

It is straightforward to check that $D_i^* \subseteq C_i^{(\infty)}$ in general and that $D_i^* = C_i^{(\infty)}$ for two–player games. That is, rationalisable strategies, in general, are a subset of the set of strategies that are left when the process of successive elimination of strongly dominated strategies is completed; while, in two–player games, they coincide with such a set.

The Nash equilibrium

This subsection introduces the most powerful, popular and controversial tool of game theory. It comes from John Nash who gave this discipline, through a number of seminal papers in the 1950s, an impetus and character that it retains. In these papers, he addresses precisely the problem of multiple rationalisable strategies by seeking to place further restrictions on the beliefs which a rational player will entertain.

In Subsection (2.2.1), I presented the so–called Harsanyi doctrine, which (together with the Aumann argument over the impossibility of agreeing to disagree) implies *consistent alignment of beliefs* (henceforward, CAB). Put informally, the latter notion means that no instrumentally rational person can expect another similarly rational person who has the same information to develop different thought processes. Or, alternatively, that no rational person expects to be surprised by another rational person.

Starting from the assumption of CAB, Nash developed his solution concept. To illustrate his argument, let us therefore suppose that players' beliefs are consistently aligned. Then each player i would want to choose the pure strategies that maximise his expected payoff, and there should be zero probability of his

choosing any strategy that does not achieve this maximum. That is,

$$\text{if} \quad \sigma_i(c_i) > 0 \quad \text{then} \quad c_i \in \underset{d_i \in C_i}{\text{argmax}} \quad u_i(\sigma_{-i}, [d_i]) \tag{2.2}$$

where $\sigma_i(c_i)$ represents the probability that i would choose c_i, and $u_i(\sigma_{-i}, [d_i])$ is i's expected payoff if he uses the pure strategy d_i while all other players behave independently according to the randomised–strategy profile σ.

A randomised–strategy profile σ is a *Nash equilibrium* iff it satisfies condition (2.2) for every player i and every c_i in C_i.[9]

Thus, a randomised–strategy profile is a Nash equilibrium if and only if no player could increase his expected payoff by unilaterally deviating from the predictions of the randomised–strategy profile. That is, σ is a Nash equilibrium iff[10]

$$u_i(\sigma) \geq u_i(\sigma_{-i}, \tau_i), \quad \forall i \in N, \quad \forall \tau_i \in \Delta(C_i). \tag{2.3}$$

Put differently, we can say that a strategy profile is a Nash equilibrium if its implementation is not inconsistent with the expectations of each player about the others' choices, or also that Nash strategies are the only rationalisable ones which, if implemented, do not contradict the expectations on which they were based. This is why they are often referred to as *self–confirming strategies* or why it can be said that this equilibrium concept requires that players' beliefs are consistently aligned.

Notice that a corollary of definition (2.3) is that Nash equilibria are formed by strategies that are best responses to each other. This reveals the connection between this equilibrium concept and CAB from another angle. If one accepts the Harsanyi doctrine and all players face the same information about the rules of the game, then one will accept that the players will draw the same inference about how rationality requires them to play. We assume common knowledge of rationality; so the players will expect that the uniquely rational way of playing will be followed. The question is: what is it? If there is one way for rational agents to play and they are all rational, then it follows that this uniquely rational way must satisfy the condition of specifying strategies which are best responses to each other. Otherwise a player, by not selecting a best response, will not be acting rationally. Thus one may not be able to see immediately what the uniquely rational way of playing is, but he can narrow the answer

[9] Cf., Nash (1951).

[10] For a proof of the equivalence between condition (2.2) and the following condition (2.3) see, for instance, Myerson (1997).

down because of his knowledge that, when there exists a uniquely rational way, then it will have to be formed by strategies that are best responses to each other; i.e., they must be in Nash equilibrium.

The thinking behind the Nash equilibrium is in many respects quite brilliant. By putting more restrictions on the beliefs hold by rational players, it arrives at a simple conclusion which corresponds to the highest degree of mutual respect of everyone's mental capacities. In more practical terms, it can furnish a set of solutions that are significantly smaller than the sets of rationalisable or iteratively undominated strategy profiles. It is thus no wonder that game theorists, as well as many social theorists, have embraced the Nash concept. Nevertheless (as many game theorists admit,[11] and as will be pointed out at the end of Chapter 3) there exist reasons for being cautious.

Subgame–perfect equilibria

The concept of subgame–perfect equilibrium was defined for extensive–form games by Selten (1965; 1975; 1978).

Let Γ^e denote an extensive–form game. For any node x in Γ^e, let $F(x)$ be the set of all nodes and branches that follow x, including the node x itself. We say that x is a *subroot* if any player who moves at x or thereafter knows that the node x has occurred. A *subgame* of Γ^e is any game which can be derived from Γ^e by deleting all nodes and branches that do not follow some subroot x and making that node x the root of the subgame.

Let Γ^e_x be a subgame of Γ^e which begins at a subroot x. If node x occurred in the play of Γ^e, then it should be common knowledge among the players who move thereafter that they were playing in the subgame beginning at x. That is, a game theorist who was modelling this game after node x could describe the commonly known structure of the situation by the extensive–form game Γ^e_x, and could try to predict players' behaviour just by analysing this game. Rational behaviour for the players in Γ^e at node x and thereafter must also appear rational when viewed within the context of the subgame Γ^e_x. Thus, Selten (1965; 1975) defined a *subgame–perfect equilibrium* of Γ^e to be any equilibrium of Γ^e which is also an equilibrium for every subgame of Γ^e. In other words, subgame–perfect equilibria require that an equilibrium should remain an equilibrium when it is restricted to a subgame, which represents a portion of the original game that would be common knowledge if it occurred.

[11] See, for instances, Binmore (1990); Kreps (1990); and Hargreaves–Heap and Varoufakis (1995).

2.3 Social dilemma games: definition and relevance

Interest in social dilemmas has grown dramatically in the past 20 years among humanists, scientists and philosophers. Social dilemmas are defined by two simple properties: (*a*) the payoff to each individual for defecting behaviour is higher than the payoff for cooperative behaviour, regardless of what the other society members do, yet (*b*) all individuals in the society receive a lower payoff if all defect than if all cooperate.

This section reviews the structure and ubiquity of social dilemma problems, outlines the specific dilemma game (i.e., the public goods problem) analysed here, and then emphasises the game–theoretic prediction in such a dilemma.

2.3.1 Introduction to the logic of social dilemmas

A social dilemma is, by definition, a situation in which each group member gets a higher payoff if he goes after his individual interest, but the whole group is better off if all group members pursue the common interest.[12]

Examples abound. During pollution alerts in some areas, residents are asked to ride bicycles or walk rather than to drive their cars. Each person is better off driving rather bicycling or walking; yet all the residents are worse off using their cars and maintaining the pollution than they would be if all bicycled or walked.

Soldiers who fight in a big battle can reasonably conclude that no matter what their comrades do they personally are better off taking no chances; yet if no–one takes chances, the result will be a rout.

Or consider the position of a wage earner who is asked to use restraints in his salary demand. Doing so will hurt him a lot; yet if all fail to exercise restraints, the result will be high inflation from which all will suffer.

Women in India may outlive their husbands and, for the vast majority who cannot work, their own old age's source of support would be their male sons. Thus, each womàn achieves the highest social payoff by having as many children as possible. Yet the resulting overpopulation makes a social security or old–age benefit system impossible, so that all the women are worse off than they would have been if they had all practised restraints in having children.

Some of these examples come from the three crucial problems of the modern world: resource depletion, pollution and overpopulation. In most societies, it is to each individual's advantage to use as much energy, to pollute as much, and to have as many children as possible. Yet the result is to exceed the "carrying

[12] For a precise definition of social dilemmas see, for instance, Dawes (1975; 1980).

capacity" of "spaceship earth",[13] an excess from which all people will eventually suffer. It is these dilemmas, which are particularly global and pressing, that have attracted the most attention among social thinkers.

Probably one of the most influential article on such a topic is Garrett Hardin's (1968) *Tragedy of the Commons.* In it (p. 1244), Hardin argues that modern humanity as the result of the ability to overpopulate and overuse resources faces a problem analogous to that faced by herdsmen using a common pasture. Hardin's situation does indeed result in a social dilemma, although not all social dilemmas have that precise form. In particular, it is a dilemma situation in which the external consequences of each herdsman who is trying to maximise his own profits are negative, and these negative consequences outweigh the positive ones to him. This brings the economic concept of *externality* to mind.[14] In Hardin's commons (as in almost all social dilemmas) the externalities are negative and greater than the individual's payoff.

2.3.2 The public goods (free–riding) problem

The public goods problem is a typical social dilemma (in the sense that it meets the two conditions stated at the beginning of this section) and it is essentially an externality problem. Individuals have scarce resources to allocate among alternative uses. Resources allocated to some uses benefit only the individual, while resources allocated to other uses benefit others besides the individual. An individual who decides to allocate resources on the basis of private costs and benefits will neglect these external benefits and allocate too few resources to those uses from which others benefit.

A typical *symmetric* public good game can be modelled in the following way. There are N identical players, each of which is endowed with m tokens. These tokens can be either contributed to a public good (and used to produce units of this good) or privately consumed. Let g_i denote the amount contributed to the public good by player i ($\forall i \in N$), with $g_i = \{0, 1, 2, ..., m\}$. Each player's earnings from private consumption is simply the amount consumed; i.e.: $(m - g_i)$. Each player's earnings from contributions to the public good is a multiple of the sum of all players' contributions; i.e.: $v \sum_{j=1}^{N} g_j$. The payoff function of

[13] Cf., Hardin (1976).

[14] "We can define an externality as being present whenever the behaviour of a person affects the situation of other persons without the explicit agreement of that person or persons" (Buchanan 1971, p. 7).

a representative individual i typically is linear and takes the following form:

$$U_i = U(g_i, \sum_{j=1}^{N} g_j) = (m - g_i) + v \sum_{j=1}^{N} g_j, \quad \forall i \in N. \tag{2.4}$$

The marginal rate of substitution of public contributions for private consumption, or marginal per capita return (MPCR), is here equal to v. Let us assume that:

$$\frac{1}{N} < v < 1. \tag{2.5}$$

Inequality (2.5) creates a dilemma situation as defined before. When all the others contribute m to the public good, player i's ($\forall i \in N$) payoff is always higher if he does not contribute anything than if he contributes m; i.e.,

$$U(0, (N-1)m) > U(m, Nm);$$

but universal contributions among the N players leads to a greater payoff than does universal defection; i.e.,

$$U(m, Nm) > U(0, 0).$$

What should one expect to happen in this simple public good game?

2.3.3 Game theoretic predictions

Given inequality (2.5), the unique dominant–strategy Nash equilibrium of the game is, for each subject, contributing nothing to the public good. Indeed, each one token contributed yields only v to its contributor (and costs him 1) no matter what the others do. However, if everyone chooses the equilibrium, the resulting outcome is less preferred by all players to that resulting from all group contributing. The equilibrium is, therefore, not Pareto efficient.[15]

Hence, the public goods (free–riding) problem can be identified as one in which all players have dominating Nash strategies that result in a non Pareto efficient equilibrium.

If Nash equilibria can be interpreted as describing how rational players should play the game, then rational individuals should expect to all do relatively badly in this game.

[15] An outcome of a game is (*weakly*) *Pareto efficient* iff there is no other outcome that would make all players better off.

Repeated public goods games

People may behave quite differently toward those with whom they expect to have a long–term relationship than toward those with whom they expect no future interaction. For instance, it becomes possible to condition what one does on what his partners have done in the previous rounds. Thus one can punish or reward the others depending on their past behaviour. Likewise one learns things about his partners from the way in which they have played, and he can exploit such learning by using early rounds of the game to develop and secure a reputation in later rounds. Therefore the analysis of repeated games promises further insights regarding the behaviour that we may expect from rational players.

(i) *The finitely repeated game*

When the game is repeated a fixed number of times (that is, when the players know in advance when the game will end), repetition does not encourage people to contribute. To understand why this happens, suppose that it is common knowledge that the game will be played for exactly 100 rounds. Then, at the last (100th) round, contributing to the public good cannot induce any further contributions by the other players (because there is no future; the last play is, after all, just a one–shot version of the game); so there is no reason for each player to contribute. Hence, all rational players should contribute 0 at the 100th round, no matter what the prior history might be. At the 99th round, the players must know that their moves will have no impact on the 100th–round moves, so they should contribute 0 on the penultimate round as well. Working backward through the game, it is straightforward to verify that the unique subgame–perfect equilibrium is to contribute 0 at every round.[16]

(ii) *The randomly–infinitely repeated game*

Suppose now that the number of times that the game will be played is a random variable, unknown to the players until the game stops. In this case, contributing behaviour can be supported in equilibrium. The key is that, whenever the players meet, they believe that there is a high probability that they will play again; so the hope of inducing future contributions by the other players can give each player an incentive to contribute.[17]

In infinitely repeated games with standard information (which represent situations in which a group of individuals face exactly the same situation infinitely

[16] See J. W. Friedman (1986).

[17] On this point see Myerson (1997, p. 309), who explains the concept for the repeated prisoner's dilemma.

often and always have complete information about each other's past behaviour), a strategy for a player is a rule for determining his move at every round as a function of the history of moves that have been used at every preceding round. One celebrated strategy for repeated games like the social dilemmas is the so–called *tit–for–tat* strategy. Under it, a player chooses to cooperate in the first round, and thereafter chooses the same move of his partners in the preceding round.

If all players follow the tit–for–tat strategy, then, in every round, the actual outcome can be that in which all cooperate if the discount factor of average pay-off assumes specific values. However, even if it is an equilibrium for all players to use the tit–for–tat strategy, such an equilibrium might not be subgame–perfect.

2.3.4 How can the public goods problem be theoretically solved?

Since the public goods problem is a pervasive phenomenon of social life, the question arises of how it is possible to deal with it.

Economic theorists have reacted to the problem by exploring institutional designs which might facilitated contributions and by designing sophisticated mechanisms for the implementation of an efficient allocation of public goods.

Designing a mechanism to foster contribution is not at all easy. The difficulties lie in finding a normative standard that would help contribution and, then, ensuring that individual incentives are aligned with the standard. It is well understood that the lack of alignment between the standard and the incentive can be a serious inhibitor to cooperation.

For years a fundamental belief was that such an alignment was impossible regardless of the normative standard. That is, it was believed that all imaginable normative standards would suffer from the problem of alignment. An intuition about both the nature and the depth of such a fundamental belief is captured by the following quote of Samuelson: "One can imagine every person in the community being indoctrinated to behave like a parametric decentralized bureaucrat who *reveals* his preferences by signalling in response to price parameters, ... to questionnaire, or to other devices. But there is still this fundamental technical difference going to the heart of the whole problem of *social* economy: by departing from his indoctrinated rules, any one person can hope to snatch some selfish benefit in a way not possible under the self–policing competitive pricing of private goods; and the 'external economies' or 'jointness of demand' intrinsic to the very concept of collective goods and governmental activities makes it impossible for the grand ensemble of optimizing equations to have that special pattern of zeros which makes *laissez–faire* competition even

theoretically possible as an analogue computer" (Samuelson 1954, p. 389).

The research world was shocked with the publication of the Groves and Ledyard's (1977) paper, in which the authors produced the outline of a process that demonstrated that the fundamental belief is wrong. In the world of theory at least, there is no logical incompatibility of purpose. The Groves–Ledyard mechanism formulates a particular allocation–taxation scheme such that individuals find it in their self–interest to reveal their true preferences for the level of public goods provided. Furthermore, the resulting level of public goods would be Pareto optimal; that is, it would be the same as if individuals had been "indoctrinated to behave like a parametric decentralized bureaucrat" described in Samuelson's quote. In other words, the Groves–Ledyard mechanism provides a set of incentives for individuals to reveal their true demand for the public goods. So, the incentives of individuals become perfectly aligned with the 'normative standard'.

Other economic theorists have suggested alternative mechanisms for enforcing an efficient allocation of public goods.[18] However, the proposed mechanisms are frequently rather complicated and difficult to implement. In his survey, Laffont writes: "... any real application will be made with methods which are crude approximations to the mechanisms obtained here ... considerations such as simplicity and stability to encourage trust, goodwill and cooperation will have to be taken into account" (Laffont 1987, p. 567).

Recently, several authors have proposed mechanisms which induce efficient contributions to the public good and seem to meet the requirements of simplicity. Varian (1994a), for instance, examined a simple two–stage game in which agents had the opportunity to subsidise the others' contributions. His mechanism relies on the concept of subgame perfection.[19] Earlier, Bagnoli and Lipman (1989) presented a simple sequential voluntary contribution game which implements the core of the economy. However, the implementation requires a rather complex and particular refinement of the Nash equilibrium.[20]

Other recent proposals deal with mechanisms in which the government tries to increase contributions to the public good by suitable tax–subsidy schemes. Such mechanisms are not purely private since they require a central authority which can enforce taxes. The relevant question is whether a government with

[18] For a survey see Laffont (1987).

[19] Varian (1994b) generalised the mechanism to other economic environments involving externalities. Guttman (1978; 1987) and Danzinger and Schnytzer (1991) considered a similar game in which individuals chose subsidy rates in the first stage, and decided about their contributions to the public good in the second stage, given the subsidy rates chosen in the first stage. A critical assessment of their analysis is provided by Althammer and Buchholz (1993).

[20] Bagnoli and Mckee (1991) and Bagnoli et al. (1992) tested this mechanism experimentally.

no information about private characteristics can design a tax–subsidy scheme which induces people to contribute more. The literature on the neutrality of lump–sum payments[21] or income taxation[22] shows that this is not a trivial task.[23] Andreoni and Bergstrom (1996) put forward an interesting model of tax–financed government subsidies to private contributions which definitely increase the equilibrium supply of public goods. Falkinger (1994) has shown that the private provision of public goods increases significantly if people value the relative size of their contributions positively. In Falkinger (1996) a simple tax–subsidy scheme is designed which induces people to take into account the relative size of their contributions in such a way that an increased or even an efficient level of public good provision is achieved as a Nash equilibrium.[24]

All the above mentioned mechanisms are desirably simple and do well in theory. It is, however, important to note that the fact that a mechanism does well in theory, does not say much about its effectiveness in the laboratory or in practice. In principle, it could well be the case that, although the Nash equilibrium in the presence of the mechanism implies an efficient provision of the public good, subjects' actual behaviour will generate significant under– (or over–) provision. For instance, in a recent paper by Chen and Plott (1996) it turns out that the performance of the Groves–Ledyard mechanism critically depends on the *size* of a so–called punishment parameter, which, according to the theory, should not affect at all the performance of the mechanism.[25]

All these institutional designs start from the assumption that individuals facing a public goods situation will behave according to the economic/game–theoretic predictions and, hence, they will always free–ride. In other words, the proponents of this literature admit the existence of the public goods problem and try theoretically to solve it by designing mechanisms that would make 'rational' people reveal their preferences for the public good.

The question of overriding importance therefore becomes: is it certainly true that there is a public goods problem? That is, are individuals fully rational and self–interested as the theory predicts or do they deviate from the predictions? To answer these questions, we need to discover what happens in a voluntary contribution context where institutional designs and public poli-

[21] Cf., Warr (1982; 1983).

[22] Cf., Bernheim (1986).

[23] See Brunner and Falkinger (1995) for a general characterisation of neutral and non–neutral taxes and subsidies in an economy with private provision of public goods.

[24] Falkinger et al. (1999) report on a series of experiments designed as a test of the practical tractability and effectiveness of the incentive mechanism proposed by Falkinger (1996).

[25] In particular, in order to implement the efficient solution as a Nash equilibrium, the mechanism only requires that the punishment parameter must be positive.

cies are absent. Experimental studies on the provision of public goods through voluntary contributions will be the topic of the next chapter.

Chapter 3

How people actually behave in dilemma game situations: a first generation of experiments

3.1 Introduction

Until the beginning of the 1970s the public goods problem (that individual incentives are at odds with group interest so that any individual who has a chance to free–ride will take it) was the conventional wisdom among economists. However this view became widely recognised as an assumption rather than a fact: there were indeed few data to affirm that the problem really existed.

The researchers' interest was therefore directed to understand to what extent people actually try to free–ride when there are incentives for doing so. Experimental methods in economics as well as in psychology provided us with important tools for dealing with such an issue, and the systematic experimental effort carried out in the 1970s and in the 1980s by various research groups has been fundamental in developing our understanding of the problem.

What emerges from these early experiments is that—contrary to the theoretical predictions—people are willing to cooperate. Generally speaking, the major findings of this vast literature are the following:

1. in one–shot trials and in the initial stages of finitely repeated trials, subjects contribute around halfway between the Pareto–efficient level and the free–riding level;

2. contributions decline with repetition; this decay is observed when subjects

know the length of the game for sure[1] as well as when they do not know;[2]

3. face–to–face communication improves the rate of contributions.

This chapter will report on some of these pioneering experiments directed to study people's behaviour in the presence of public goods. Before analysing how the experimental method has been used to provide insights into this topic, it seems appropriate to discuss some general methodological issues concerning experimentation in economics.

3.2 Experimental methods in economics

For many years it was a widely shared belief that economics is not an experimental discipline. Friedman, for instance, affirms: "Unfortunately, we can seldom test particular predictions in the social sciences by experiments explicitly designed to eliminate what are judged to be the most important disturbing influences" (M. Friedman 1953, p. 10). And, quite a few years later, Samuelson and Nordhaus declare: "One possible way of figuring out economics laws... is by *controlled experiments*... Economists [unfortunately]... cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors" (Samuelson and Nordhaus 1985, p. 8).

Nevertheless, when Charles Plott, in his Presidential Address to the VI Annual Meeting of the Southern Economic Association in 1990, raised the question "Will economics become an experimental science?" he confidently predicted that it would.

Economists have been employing the experimental method for at least 60 years. However, for the first part of this period, this method had only limited significance in empirical economics: experiments were quite rare and often conducted in a rather informal manner. During the last 25 years the picture has changed considerably. The use of experimental methods to address economic questions has grown rapidly and experimentation has acquired a significant degree of recognition as a legitimate branch of empirical enquiry, relevant to economic discourse.[3] These days, there are many economists who speak enthusiastically about experiments and their contributions to economics. For instance,

[1] Cf., Isaac et al. (1984); and Isaac and Walker (1988).

[2] Cf., Kim and Walker (1984); and Isaac et al. (1985).

[3] There are numerous clear signs of this: articles reporting experimental research are now frequently published in leading international journals such as *Econometrica*, the *American Economic Review* and the *Economic Journal*; the *Journal of Economic Literature* established a separate bibliographic category (*Experimental Economic Methods*) devoted to the classification of experimental works in the late 1980s; and a new journal (the *Journal of Experimental Economics*) entirely devoted to experiments began publication in 1998.

two leading experimentalists maintain: "Since the mid–1970s this kind of work [experimentation] has been transformed from a seldom encountered curiosity to a small but well–established and growing part of the economic literature" (Roth 1987, p. 1), and "... as currently practised, economics is ideally suited for experimental investigation" (Hey 1991, p. 15).

This enthusiasm is to a large extent at odds with the sceptical tones of the opening quotations. Given the large amount of experimental work produced currently, the view that experiments are rarely possible in economics seems hard to sustain. Hence, if taken simply as suggestions that one cannot run experiments in economics, the sceptics' comments appear to contrast with the current events. There is, however, another different perspective of reading these critical comments, i.e., that experimental procedures may not provide very meaningful data relevant to the discussion of economic enquiry.

The latter is certainly a more plausible interpretation of the statement of Friedman who, at the time of his *Essay in Positive Economics* (1953), was well aware that economists had been using experimental methods. In fact, Friedman was co-author of a review of one of the first experiments reported in the economics literature: that of Thurstone (1931). In such a review we read: "It is questionable whether a subject in so artificial an experimental situation could know what choices he would make in an economic situation; not knowing it is almost inevitable that he would... systematize his answers in such a way as to produce plausible but spurious results" (Wallis and Friedman 1942, p. 179).

Arguments to the effect that experimental results may be 'spurious' because of the 'artificial' context in which they are generated should be familiar to anyone who has presented experimental results to an audience of general economists. Indeed, Loomes notes that one of the questions most frequently posed to experimentalists is: "Can you *really* take the observed behaviour of groups of volunteers spending short periods of time in carefully controlled environments and draw any meaningful conclusions about the way things work in the outside world?" (Loomes 1991, p. 29).

This question exemplifies a scepticism towards experimental research in economics which persists in spite of the now widespread use of the method and its new prominence in mainstream dialogues.

I will examine what basis there might be for this scepticism later in this section, after a brief discussion of the reasons why laboratory experiments have become such an important source of data for economists.

3.2.1 Why laboratory experiments? Their main purposes and advantages

Several aims can be pursued with experimentation. A review of them is in order.[4]

A primary purpose of experiments is to discover empirical regularities in areas where no theory exists. An important example of 'theory searching experiments' is provided by market experiments. Much of the literature in this area is concerned with the existence and the properties of the equilibrium but not with its achievement. Some kind of meta–theory would be required to show how a set of consumers and producers actually reach the equilibrium. Experiments have filled this theoretical gap by revealing that market processes converge to the equilibrium with a speed which depends on the institutional setting.[5] Smith (1982) calls these experiments *heuristic.*

In other areas, by contrast, several competing theories offer different predictions. In this case, the role of laboratory work is to delineate the range of applicability for each theory. For example, Fiorina and Plott (1978) studied committee decisions in the laboratory and found that only a few of the sixteen models and variants considered were at all consistent with the data. Finally, there exist areas for which only one model is applicable. Here, experimentation can demonstrate whether there are any conditions under which the theory can account for the data and, if so, it can test theory for robustness. Smith (1982) refers to the last two types of experiments as *boundary* and to sets of experiments directed to establish definitive broad laws of behaviour as *nomothetic.*

A further purpose that can be fulfilled by experiments is to advise policy–makers in settings lacking theoretical predictions. For example, Grether and Plott (1984) used experiments to provide evidence in an antitrust case. Often the question is which institution yields the most efficient allocation in an economic or political market. Recently, economists and policy–makers have found it useful to study new institutions in the laboratory before introducing them in the field.

Some experimental economists believe that one purpose of running experiments is to *test a theory.* Most theorists oppose this opinion in that they see prediction as the primary end of theorising. According to them, a theory is formally valid if it is internal consistent (that is, it does not lead to statements that contradict each other) and if the conclusions are provable from the assumptions. This is certainly the view of Friedman (one of the above sceptics) who

[4] Cf., Plott (1982; 1987).

[5] On this topic see, for instance, Hey (1991).

grants that a theory is of direct interest only to the extent that its conclusions provide good approximations of actual behaviour even when its assumptions are not precisely satisfied.[6] The debate on this topic among economists (both experimentalists and not) is still open and, given its complexity, it cannot be adequately treated here.[7]

As far as the advantages of the experimental method are concerned, the most important one is that it allows the researcher to have *control* over the environment. The institutions are easy to control in a laboratory but not in the outside world; hence, it is easier to evaluate the effects of a change in institutions in an experimental setting. The ideal procedure in an experiment lets the researcher to measure the effects of one of the independent (exogenous) variables on the dependent variable by varying this independent variable, while keeping the others constant. In this way, casual relationships can be traced and alternatives theories and policies can be evaluated and compared. Although sometimes experimenters do not succeed in reaching this ideal,[8] they usually do a better job than field studies where there is virtually no control over important variables.

Of course there exist aspects which are simply uncontrollable. Examples are subjects' personal idiosyncrasies and (usually) expectations. Random allocation of subjects to different treatments takes care that treatment effects are not affected systematically by such uncontrolled aspects. Besides randomisation, in an experimental setting it is also possible to elicit and measure expectations (or idiosyncrasies) which are usually considered unobservable in field studies (even if economists—unlike psychologists—are not so keen on obtaining such information and measurements). In this study, I will elicit expectations that subjects have regarding the others' behaviour and (by means of a questionnaire) I will try to understand subjects' attitude towards other people.

A further advantage of experimentation is that its data are *replicable*.[9] Field data are generated from events that occurred at a specific time in a specific place. Due to the continually changing nature of these settings, it is very difficult for other researchers to replicate a field data set, therefore making it difficult to verify the accuracy of both the data and the findings. Since laboratory data are generated in controlled conditions, it is easier to reproduce an

[6] Cf., Friedman (1953, p. 23).

[7] For references (besides the cited Friedman's (1953) essay) see also Koopmans (1957); Friedman and Sunder (1994); Lipsey and Crystal (1995); and Starmer (1996).

[8] Examples of experiments which can be criticised for 'lack of control' will be provided later in this chapter.

[9] Cf., Davis and Holt (1993, p. 14).

experiment and replicate its results. Furthermore, if it is suspected that a data set contains a lot of noise or if the data are insufficient to draw a firm conclusion, one may re–run the experiment and collect new data. With naturally occurring phenomena it may take considerable time to acquire a different data set of the desired size.

3.2.2 Realism and controlled economic experiments

As pointed out above, a usual objection raised against experiments is that they lack realism. The argument is that subjects behave differently in an 'artificial' laboratory setting than in the real world, and that—as a consequence—the data obtained in experiments are meaningless.

Actually, an effective design is often very simple compared to reality. There are practical, financial and ethical considerations which render certain sorts of realism impractical. But—as many authors affirm—it is futile to try to replicate in the laboratory all features and complexities of some naturally occurring phenomena.[10] There exists indeed a trade–off between realism and control. If one wants to make the laboratory environment close to reality, then the experiment will become so complex that one will find it difficult (or even impossible) to clear up causes and effects and so the experimenter will lose control over some relevant variables. If one wants to have control over the important variables, then one must find the simplest laboratory environment which incorporates the key interesting aspects of the real world.

Plott argues that many early experimentalists mistakenly believed that "the only effective way to create an experiment would be to mirror in every detail, to simulate, so to speak, some ongoing natural process" (Plott 1991, p. 906). But a good part of the rationale for experimenting, in Plott's view, is that it allows us to investigate the relationships hypothesised in a theory while abstracting from other factors which, even if at work in the broader social setting, are not part of the theory being investigated. In other words, in so far as is possible, the experimenter must *deliberately* omit these other factors. Therefore: "Once models, as opposed to economies, became the focus of research the simplicity of an experiment and perhaps even the absence of features of more complicated economies became an asset. The experiment should be judged by the lessons it teaches about theory and not by its similarity with what nature might happened to have created" (Plott 1991, *ibidem*).

The issue therefore is not how to replicate a real world decision setting in every detail but how to create an appropriate abstract setting which isolates

[10] Cf., Plott (1991); Friedman and Sunder (1994); and Starmer (1996).

all the interesting elements of the theory under consideration. In other words, the question is: how can we tell when an experimental environment has been suitably *controlled* to allow an appropriate test of a given hypothesis? I will illustrate the point in relation to a particular class of experiments.

A large number of experiments have been concerned with investigating the behaviour of *microeconomic systems.* Smith (1982; 1989) argues that any microeconomic system consists of two elements: *environment* and *institution.* In Smith's terminology, the environment refers to: the set of agents participating in the system (including their individual characteristics such as their preferences); the goods existing in the economy; the production technology; and the initial resource endowments. The institution governs the interactions of the agents: it can be thought of as a set of rules which define how agents may or may not interact and it specifies the form that any such interaction must take (including the language through which messages may be transmitted). The interplay of the environment and institution generates *behaviour* which refers to all the observable outcomes of the system: the agents' actions and the emergent outcomes, for example. In experiments with microeconomic systems, the experimenter seeks to create and (in most cases) to manipulate institution and environment with the intention of investigating the relationship between their characteristics and the behaviour of the system.

A great number of experiments can be interpreted in such a way. For instance, let us consider the public goods experiment described in Chapter 2, Subsection (2.3.2). The individuals in the experiment, their preferences and their initial endowment constitute the environment. The rules of the experiment—how contributions can be made and how they are redistributed to participants—define the institution.[11]

Is the behaviour observed in these experimentally generated microeconomies just an artifact of the laboratory system or is it of more general significance?

In a sequence of closely related papers, Smith (1976; 1980; 1982) and Wilde (1980) have argued that, *given certain assumptions*, the laboratory microeconomies can be regarded as real (small–scale) microeconomic system; the agents' behaviour can be thought of as a real economic behaviour and, therefore, the observed behaviour is suitable as a test for economic theories. These assumptions are related to the reward structure of the experiments and are the following.

[11] In Subsection (2.3.2) the institution is represented by the *voluntary contribution mechanism* (without communication). In it, each individual is told to contribute an amount of his initial endowment privately and without any information about what the others are doing. The level of public good provided then equals that producible with the total endowments contributed.

Nonsatiation (or *monotonicity*) which requires that subjects prefer more money to less, and that this desire does not become satiated in the course of the experiment. *Saliency* which requires that the reward received by a subject depends on his action (and possibly on the actions of the other agents).[12] *Dominance* which requires that changes in subjects' utility from the experiment come predominantly from the reward medium and that the effects of other costs (or benefits) on their utility are negligible.

A further condition introduced by Smith to neutralise motivational factors that may disturb the reward structure induced by the experimenter is *privacy*. This condition requires that each subject receives information only on his own reward, and it is intended to guard against motivations arising from interpersonal considerations. However, some kinds of experiments (among which public goods experiments) are expressly directed to investigate if (and to what extent) subjects care about the rewards earned by the others; in these cases, the privacy condition is intentionally excluded by the experimental procedures.

According to Smith (1982, pp. 931–935), if nonsatiation and saliency are satisfied, we are entitled to interpret the behaviour observed in an experimental microeconomy as maximising behaviour which is appropriately tied to the institutional context of the experiment.

When also dominance holds, the subjects are maximising utility functions which depend only on the reward medium designed by the experimenter, who has, hence, achieved control over subjects' preferences in the sense that he has induced known preferences on the subjects.[13]

This argument constitutes a challenge to those who argue that the laboratory is an artificial environment and that, consequently, the results may be spurious because they do not capture all the complexities of the real world economic environment. The Smith/Wilde position (like that of Plott) shifts the focus from comparison between laboratory and real world to comparison between laboratory and theory. The microeconomy created in laboratory is much more close to the economic theory than does the real economy. Hence, if the theory fails in this simple context, there are good reasons for expecting that its predictions will also fail in a more complex natural environment.

There seems to be great merit in this argument. Nevertheless, its practical significance is severely limited by two problems.[14] The first is that it will never be possible to establish whether the conditions have been satisfied in a given

[12] For example, a fixed flat payment is not salient because it does not depend on the actions chosen by the subjects in the laboratory.

[13] Cf., Friedman and Sunder (1994, p. 13).

[14] Cf., Cross (1980).

experiment. For instance, to determine if dominance has held, one needs to ask whether the experimental rewards were sufficiently large to outweigh other possible influences. Some economists believe that experiments must use 'adequate' incentives before the results can be taken seriously.[15] The problem is to know what 'adequate' means and there is no agreement among experimentalists on this point. Some empirical work suggests that incentives do affect behaviour in the sense that subjects take their task more seriously with incentives,[16] while other work reports no effect of incentives on behaviour.[17]

The dominance condition seems to be especially problematic in public goods and bargaining games. In these areas, indeed, subjects' choices seem to be affected by concerns about the payoffs for others or concerns for being treated fairly. One may continue to defend the view that dominance and control over preferences will result if incentives are sufficiently high, but a problem of such a defence is that it cannot be rejected.

In this study, subjects will be motivated with proper incentives as much as possible. However, in the experiments reported in Chapters 6 and 7, the test of specific behavioural hypotheses will require to assess the concerns that subjects have towards the payoffs to others; therefore, in these experiments, each subject will be informed not only about his own payoffs (as privacy requires) but also about the others' payoffs.

Similarly, for saliency and nonsatiation to hold, it is important that subjects properly understand the experiment; in particular, how the reward medium is related to their actions. And, even if they understand what has been described to them, there might be questions regarding whether subjects trust the experimenter. There are strategies that experimentalists can (and often) adopt for reducing the possibility of misunderstanding.[18] There may also be possibilities for attempting to assess subjects' understanding or to explore their trust in the experimenter.[19] But while these procedures may reduce the scope for scepticism, they are unlikely to eliminate the possibility of doubt.

The second practical limitation of the Smith/Wilde argument concerns the *generalisability* of the experimental results. In particular, Cross (1980) proposes two reasons why behaviour in the laboratory might be different from real world behaviour. First, real world behaviour may be a product of learning and

[15] See, for example, Hey (1991); and Binmore (1993).

[16] See, for instance, Grether (1992); Offerman and Schram (1993); Smith and Walker (1993); and Harrison (1994).

[17] Cf., Camerer (1995).

[18] For instance, opting for simple designs, ensuring that the instructions are clear, allowing time for subjects to become familiar with the experimental setting.

[19] Including, for instance, 'test' questions or post–experimental questionnaire.

adaptation and, if laboratories do not embody analogous mechanisms (or allow them sufficient time to operate effectively), there will always be systematically differences between the laboratory environment and the naturally occurring behaviour. Second, behaviour may be sensitive to context since human beings interpret their surroundings to decide which modes of behaviour are appropriate to their environment. Smith argues that attacks along these lines do not subvert the role of experimentation in theory testing. His argument seems to be that if economists believe that evolution, learning, context, or whatever else are important factors in determining outcomes, then those factors should be included in economic theories.

By concluding, nobody can deny that today the experimental method is widely used among economists and that it has acquired some prominence in mainstream issues. With experimentation several goals can be pursued and this method exhibits some advantages in comparison with field studies. However, due to the puzzles arising when experiments attempt to confront theory with reality, no experimenter would argue that experiments should replace empirical research. On the contrary, each method can make up for the deficiencies of the other.

After this brief analysis of the advantages and disadvantages, aims and limits of the experimental method, let us come back to the main topic of this chapter and let us see how economists (and psychologists) have used experiments to provide evidence for the public goods problem.

3.3 First attempts to estimate demand for public goods: Bohm's experiment

One of the earliest attempts to discover experimentally if there is a public goods problem is attributed by Ledyard (1995, p. 122) to Bohm (1972).

In his paper, Bohm describes "a test involving five different approaches to estimating the demand for a public good". The test was used for a TV–show not yet show to the public but available on closed–circuit TV for representative samples of the Stockholm population if aggregate willingness to pay exceeded costs. Six groups of different size (randomly drawn from a sample of 605 persons from the age 20 to 70 of the population of Stockholm) were confronted with different demand elicitation procedures, i.e., different payment consequences in case the good were to be produced. In particular, the first five groups received the following instructions.

"Try to estimate in money terms how much you find it worth at a maximum to watch this half–hour program in this room in a little while, i.e., what is the largest sum you are willing to pay to watch it. If the sum of the stated amounts of all the participants covers the cost (Kr. 500) of showing the program on closed–circuit TV, the program will be shown; and you will have to pay:
(to group I) the amount you have stated,
(to group II) some percentage (as explained) of the amount you have stated,
(to group III) either the amount you have stated or a percentage (as explained) of this amount, or Kr. 5 or nothing, to be determined later by a lottery you can witness,
(to group IV) Kr. 5,
(to group V) nothing. In this case the participants were informed that the costs were to be paid by the SR,[20] i.e., the taxpayers in general.

Counter–strategic arguments[21] were added to instructions I, II, IV and V. The subjects in group VI, who received instructions which differed from the instructions to the first five groups, were simply asked how much they found the program to be worth at a maximum. In a second round, these people were asked to give their highest bids for a seat to watch the program and were told that the 10 highest bidders out of an alleged group of some 100 persons were to pay the amount they had bid and see the program" (Bohm 1972, p. 119).

The purpose of such a design was to test whether (as predicted by the theory) group I would understate their willingness to pay, while groups IV and V would overstate.

Data analysis showed that no significant differences (at the 5 percent level) existed between any pair of instructions I to V. This leaded Bohm to conclude that "the well–known risk for misrepresentation of preferences in this context may have been exaggerated" and people may be willing to contribute to the public good even if their own self–interest runs counter.

In Ledyard (1995), the approach used by Bohm is criticised for "lack of control".[22] In particular, Ledyard refers to three aspects of the design which suggest a lack of control. First, it did not have any known individual willingness–to–pay with which to compare the stated willingness–to–pay. Second, it misrepresented the true situation to the subjects in order to study the effect of

[20] The Swedish Radio–TV broadcasting company.

[21] Presenting a set of counter–strategic arguments in order to spell out all relevant arguments and pieces of information is, according to Bohm, one of the requirements which an empirical test of alternatives approaches must meet in order to be effective.

[22] Cf., Ledyard (1995, pp. 124–125).

large group,[23] and so even if the experimenter might hope that the subjects believed that the group was large, control might have been lost. Third, it includes counter–strategic arguments whose use in experiments is clearly controversial.

Besides Ledyard's critics, Bohm's study was, at his time, an important step towards the understanding of individuals' voluntary behaviour in public goods environments.[24]

3.4 Psychological approach to social dilemmas

While economists were putting considerable effort to get their experiments under control, psychologists were independently studying social dilemmas in their laboratory. As emphasised before, for experimental economists saliency is an essential and self–evident precept; in order to tie the experiment to the relevant theory, they take great pains to establish a clear incentive structure within an institutional framework. Most psychologists, on the contrary, feel no necessity to offer salient rewards; they believe that the admonition to subjects to 'do their best' is acceptable.

As an example of the type of experiments carried out in psychology, let us consider the study of Dawes et al. (1977).

In their experiment, 40 eight–person groups were created.[25] Each subject in each group had to decide whether to mark an X or a 0 on a card in private. Each knew that marking a 0 he would earn \$2.50 with no fine to anyone, and that marking a X he would earn \$12.00 with a fine of \$1.50 to all group members (including himself). Thus each player had an incentive to defect (i.e., to choose X) but, if all defected, no one received anything. Two payoff conditions were included in the experiment. The loss condition in which payoff to a cooperator was reduced by \$1.50 for every defector in the group; and the no–loss condition in which cooperators' negative payoffs were truncated to zero. Subjects were presented the payoffs in the form shown in Table 3.1.

One of the aims of the experimenters was to investigate the effects of various aspects of face–to–face communication on cooperation rates.[26]

[23] "The subjects were given the impression that there were many groups of the same size simultaneously being asked the same questions on other rooms elsewhere in the broadcasting company" (Bohm 1972, p. 118).

[24] Ledyard himself expresses a similar judgement (Ledyard 1995, p. 125).

[25] Sometimes less people in each scheduled group showed up so that, rather than the anticipated 320, only 284 subjects were used.

[26] A different type of communication is that in which people transmit messages by computers. At the time of Dawes et al.'s experiment, no computerised experiment had been conducted. The first public goods experiment which was run by using Plato computer system is, at the best of my knowledge, Isaac et al.'s (1984) experiment.

Table 3.1: Dawes et al.'s (1977) experiment: Payoff matrix.

Payoff to X	Number X	choosing 0	Payoff to 0
	Loss condition		
–	0	8	2.50
10.50	1	7	1.00
9.00	2	6	−0.50
7.50	3	5	−2.00
6.00	4	4	−3.50
4.50	5	3	−5.00
3.00	6	2	−6.50
1.50	7	1	−8.00
.00	8	0	–
	No–loss condition		
–	0	8	2.50
10.50	1	7	1.00
9.00	2	6	0
7.50	3	5	0
6.00	4	4	0
4.50	5	3	0
3.00	6	2	0
1.50	7	1	0
.00	8	0	–

For this reason, four communication conditions were included in the design. The no–communication condition, in which groups worked silently on an unrelated topic before making their decision in the game. The irrelevant–communication condition, in which subjects discussed the same unrelated topic for 10 minutes but were not allowed to discuss the group dilemma decision. The relevant–communication condition, in which groups discussed the dilemma situation before making their decisions, and the relevant–communication–plus–vote condition in which groups ended their discussion with a roll call–non binding declaration of intended decision.

Results of the experiment are displayed in Table 3.2, which shows the average proportion of defectors in each of the eight conditions.

Three main conclusions are drawn by Dawes et al. from these data. The first is that the effect of face–to–face communication on cooperation is highly significant: without communication and with irrelevant communication cooperation

Table 3.2: Dawes et al.'s (1977) experiment: Proportion of subjects defecting.

	Condition			
Condition	No Communication	Irrelevant Communication	Unrestricted Communication	Communication Plus Vote
Loss	.73	.65	.26	.16
No Loss	.67	.70	.30	.42

rates were only 30% and 32% respectively, while with relevant communication and with communication plus commitment they increased to 72% and 71%. The second conclusion is that commitment made no difference: "the structured communication with the vote did not elicit any more cooperation than did the unstructured communication (73% versus 72% on average), *despite the fact that every subject in the structured communication condition announced an intention to cooperate*".[27] The last conclusion is that the no–loss condition had no effect: "the possible loss manipulation was not only ineffective in explaining differential cooperation, it was ineffective in eliciting differential predictions about others' behaviour as well".[28]

Let us consider each of these conclusions in turn. Relevant talking matters a lot. This is not surprising. A more informative result would have explained what it is about face–to–face communication that leads to more cooperation. Although four types of communication were used, the data provide little information with reference to *why* when people can communicate, they cooperate more than when they cannot. Letting subjects talking for 10 minutes in an *uncontrolled* setting introduces unintended effects and a lot of contamination in the design.[29]

Also the finding that the communication–plus–vote condition did not increase cooperation in comparison with the relevant–communication condition is not so amazing (contrary to what the authors believe) if one thinks that the commitment did not arise spontaneously from the group process but was forced by the experimenters. Moreover the promise to cooperate is the only reasonable statement to make no matter what one's intentions. Thus, that each subject would have announced cooperation could be expected.

Finally, as for the lack of impact of the no–loss condition, Ledyard (1995) argues that such a result is due to the existence in it of two countervailing effects

[27] Dawes et al. (1977, p. 5).
[28] Dawes et al. (1977, *ibidem*).
[29] Cf., Ledyard (1995, p. 129).

(one which induces more defection *ceteris paribus*, and the other which induces less defection *ceteris paribus*) that could have easily cancelled each other. In addition, Ledyard shows that the direction of the effect may change according to the subjects' expectations about their group members' behaviour. "The incentive effects of the no–loss treatment are" therefore "complex and out of control" (Ledyard 1995, p. 128).

3.5 First systematic approaches to the public goods problem

At the same time as psychologists, but independently from them, Marwell and Ames were experimentally investigating the predictive power of the free–rider hypothesis regarding the provision of public goods by groups. In a series of three related papers,[30] besides testing the free–rider hypothesis, they studied the effects on contributions of several independent variables: group size, distribution of interest, distribution of resources in the group, provision points, strength of induced preferences, experience of subjects, divisibility of the public good and training of the subjects.

As an example of the experimental design implemented by Marwell and Ames, let us consider the first of their three papers.

In this experiment, 256 high school students between the ages of 15 and 17 were recruited from a sample of all homes with telephone in the Madison and Wisconsin area.[31] The study was performed in a 'natural setting', in that all contact with the subjects was by telephone and mail and subjects remained in their normal environments throughout all the experiment.

Subjects were provided with a given amount of resources (in the form of tokens) which they had to invest in either an 'individual exchange' or a 'group exchange'. The individual exchange returned a fixed amount (i.e., 1 cent) for each unit of resources invested, regardless of the other group members' behaviour. The group exchange paid its cash earnings to all members of the group by a preset formula, regardless of who invested. Thus, the subject received a share of the return on his own investment in the group exchange (if any), and also the same share of the return on the investments of each other group member. The payoff table given to the subjects for a large group with

[30] Marwell and Ames (1979; 1980; 1981).

[31] "High school–age students were selected for study because we felt that the amount of money at stake in their decision (about $5.00) would be most meaningful to young people and that at the same time these subjects would be old enough to understand the investment decision they had to make" (Marwell and Ames 1979, p. 1341).

unequal benefits (designed blue and green) and unequal resources is provided in Table 3.3.[32]

Table 3.3: Marwell and Ames' (1979) experiment: Payoff from group exchange in large, unequal–interest, unequal–resource groups.

		How much money you get if you are ($)	
If the total tokens invested in the group exchange by all group members is	Tctal money earned by the group	*Blue* ($2\frac{1}{4}c$ of each group dollar)	*Green* ($9/10c$ of each group dollar)
Between			
0–1,999	0	0	0
2,000–3,999	14.00	.32	.13
4,000–5,999	32.00	.72	.29
6,000–7,999	54.00	1.22	.49
8,000–9,999	320.00	7.20	2.93
10,000–11,999	350.00	7.88	3.21
12,000–13,999	390.00	8.78	3.57
14,000–15,999	420.00	9.45	3.85
16,000–17,999	440.00	9.90	4.03
18,000	450.00	10.13	4.12

As can be seen, while returns to the group exchange are near zero for the first 7.999 tokens invested, at the investment of 8.000 tokens the return on all invested tokens increases dramatically. At this point, which the authors call *provision point*, the group exchange returns approximately 3.8 times as much as the individual exchange for every token invested. Under these circumstances, all group members would be better off if all their resources were invested in the group exchange. On the other hand, each individual would be best off if he invested in the individual exchange while everyone else invested in the group exchange (i.e., if he free–rides on all others' investments).

Notice that the payoff structure proposed by Marwell and Ames generates multiple Nash equilibria: one at which nobody contributes (the *strong* free–rider hypothesis) and several others where everyone contributes only partially. Thus, not contributing is not longer a dominant strategy and (in case of equal distribution) contributing 44% on average is a natural focal point.

Besides testing the free–rider hypothesis, the authors were interested in analysing the effect on contributions of three independent variables: group size,

[32] What large group, unequal benefits, and unequal resources mean is discussed below.

distribution of interest, and distribution of resources.

In order to vary the group size, they informed half their subjects that their group contained 80 rather then 4 persons. "However, no individual was actually a member of a group of 80 persons. All group contained just four real subjects".[33]

In order to vary the perceived distribution of interest, they implemented two different conditions: in the 'equal' condition, all group members received identical shares of the cash produced by investment in the group exchange; in the 'unequal' condition one–fourth of all group members (the *blue* subjects in Table 3.3) received a share of the group exchange approximately 2.5 larger than the share of the other group members (the *green* subjects of Table 3.3). Resources were also either equal or unequal within groups, and followed the same proportions as the distribution of interest.

The main finding of this study was that subjects invested much more in the public good than would be predicted by the strong free–rider hypothesis. A typical group invested 57% of its resources in the public good, 28% more than needed to reach the provision point.

A second result was that the rate of contribution was less if initial endowments were unequal.

Since the presence of a provision point could have been crucial in determining the high level of contribution, in a later study, Marwell and Ames (1980) removed such a point, and designed a payoff structure in which the returns to investments in the group exchange (rather than increasing precipitously at a given point) were kept in simple proportion, regardless of the total amount invested.

The general level of contribution, after this change, replicated that of their previous research: approximately 51% of the available tokens were invested in the group exchange.

The work of Marwell and Ames provided, hence, stark and clean evidence against the standard economic predictions: in their experiments, subjects contributed and did not all free–ride.

There is, however, a note of criticism which can be made against their design and which reminds that made in Section (3.3) on Bohm's research. The critic concerns the deception that Marwell and Ames practised on their subjects regarding the group size. It is well known (and it has been claimed by many experimentalists)[34] that honesty in procedures is fundamental if the data are to

[33] Marwell and Ames (1979, p. 1345).

[34] For instance, Hey (1991) and Ledyard (1995).

be valid. Any deception can be, in fact, discovered and contaminate a subject pool.

3.6 Reaction of the economists: some first responses

Marwell and Ames studies provoked the reaction of the experimental economists who had been focusing on markets and who felt sure that the work of sociologists had to be inaccurate. In particular, two researches were carried out in direct response to Marwell and Ames: one by Kim and Walker (1984) and the other by Isaac, McCue and Plott (1985). Purpose of both these authors was to show that Marwell and Ames were wrong, and "to explore the behavior of groups within a set of conditions where we expected the traditional model would work with reasonable accuracy".[35]

The main difference with the Marwell and Ames experiment was the introduction of *repetition*: in both the Isaac and al. and the Kim and Walker experiments, subjects faced the same decision problem for a series of periods rather than just making their decision once.

As for the experiment of Isaac and al., it consisted of 9 experimental sessions[36] with 10 participants in each session (except in sessions 4 and 9).

Subjects received a participation–fee of \$5. Half of them were assigned to the so–called 'high' payoff condition and the other half were assigned to the 'low' payoff condition. Subjects were given a table which indicated both their marginal payoff and total payoff at each level of the public good from 0 to 40. The functions which generated these marginal payoffs were $\$.44 - 0.011q$ for the high types and $\$.276 - 0.008q$ for the low types (where q is the amount of public good actually chosen). Given this environment, the social optimum (which maximises total payoff) is at $q = 23$ or 24, while the Nash equilibrium is at $q = 0$ and it is a single–period dominant strategy for both types.

Data analysis from this study showed that average contribution rates decayed from 38% of the efficient contributions level in the initial period to 9% in the final period. Therefore, although in initial decision periods contribution rates resembled those observed by Marwell and Ames, they declined substantially with repetition.

With a similar design, Kim and Walker (1984) found contributions provided 41% of the maximal group payoff in the first period and decayed to 11% by the third period. Their study, however, can be strongly criticised because they de-

[35] Isaac at al. (1985, p. 51).

[36] Two of these sessions were, however, excluded from the analysis on the basis of aberrant, 'irrational' participants' decisions.

ceived their subjects. In fact, although Kim and Walker were extremely careful to try to eliminate factors of earlier studies which might have invalidated the design (factors that, according to them, involved a loss of control by the experimenter), they misled their subjects hoping that they would think they were in 100–person groups. Whether the subjects believed that or not is unknowable.

Initial laboratory examination of the free–ride problem provided, hence, quite different results. On the one hand, Dawes et al. and Marwell and Ames found far less free–riding than predicted by economists. On the other hand, Isaac and al. reported nearly complete free–riding particularly in the terminal periods of multiperiod sessions.

Isaac, Walker and Thomas (1984) tried to reconcile and identify the reasons for the divergence in contribution rates reported in these earlier investigations. Consequently, their experiment included a variety of the design features that were varied across previous experiments, including repetition, group size, marginal payoff, and experience levels.

Here, each participant knew that there would be 10 decision periods and that his endowment and payoff would remain constant over all repetitions.

As for the group size and the marginal payoff, since it was challenging to change N (the group size) keeping constant the incentives between group and self interest, it is worth considering how Isaac at al. dealt with this problem. Algebrically, the payoffs in this experiment were $U_i = p(m - g_i) + ay/N$, where $y = \sum_{j=1}^{N} g_j$ and the meaning of each other term is as in Eq. (2.4). The marginal rate of substitution, M, of the private for the public good is a/pN. The marginal group return, computed from $\sum U_i = p(Nm - \sum g_i) + a \sum g_i$, is a/p. If one increases N and nothing else, then M decreases and, hence, the incentives for individual interest increase relative to the incentives for the group interest. If one increases N but keeps M constant by increasing a, then the incentives for the group interest increase relative to the incentives for individual interest. As stated above, it seems impossible to vary the group size N without changing the incentives between group and self interest. The authors solved this issue by choosing a 2×2 design with $N = 4$ or 10 and $M = 0.30$ or 0.75; always $p = 1$. Thus, one has $a = NM$ and four parameter choices (N, M, a); i.e., (4, 0.3, 1.2), (4, 0.75, 3), (10, 0.3, 3), and (10, 0.75, 7.5). These allow the experimenters to compare a change in N keeping M constant and to compare a change in N keeping $a = 3$ constant.

Finally, as far as experience is concerned, such a parameter was measured as previous participation in similar experimental sessions.

Even if the sessions differed according to the values of these variables, all

shared the characteristic that complete free–riding (i.e., a contribution rate of zero) is the unique Nash equilibrium, and that the joint income is maximised when all participants contribute all tokens to the group exchange.

Table 3.4 provides the percentage contribution data from this experiment.

Table 3.4: Isaac et al.'s (1984) experiment: Percentage contribution data.

	Period										
	1	2	3	4	5	6	7	8	9	10	*Aver.*
All	51.1	47.2	44.1	47.4	46.7	38.1	40.6	35.2	35.8	37.3	42.4
M=0.3	43	35	28	32	26	25	20	17	20	17	26
M=0.75	60	59	60	63	67	51	61	53	52	57	58
Inexperienc.	53	53	45	50	55	43	50	41	39	44	47
Experienc.	49	41	43	45	38	33	31	30	33	30	37
$N = 4$	50	50	38	40	38	30	36	32	38	30	38
$N = 10$	56	50	40	41	41	34	32	33	37	35	40

As can be seen, the average percentage contribution across all treatments is 42%, and the average across first periods is 51%. These look very much like Dawes et al. and Marwell and Ames. However, the variance is high: contribution rates ranged from 0% (period 8 with $M = 0.3$, $N = 4$, experienced subjects) to 83% (period 5 with $M = 0.75$, $N = 4$, inexperienced subjects). Hence, something more than just 40–60 percent contribution is going on.

Three more conclusions derive from this experiment's data. First, increasing M from 0.3 to 0.75 increases contribution levels in all cases. Second, experience matters since inexperienced subjects contribute more. Finally, repetition decreases and group size increases contributions for low M ($= 0.3$) but neither seem to have an effect if $M = 0.75$.

3.7 An important final question: Is it really the theory that is wrong?

Experimental evidence from the pioneering laboratory work described in this chapter shows that people in the presence of public goods do not behave as predicted by the traditional theory: they do not free–ride all the time, but voluntarily provide public goods in one shot trials and in the initial stages of repeated trials. Hence, the question worth reflecting upon is: do we face bad experiments or a wrong theory?

Undoubtedly there are difficulties in doing experimental research in public goods and, except the Isaac, Walker and Thomas experiment (whose design was carefully thought out and confounding variables were kept under control), all the other examined designs exhibit some uncontrolled features which expose them to criticism.[37]

Nevertheless, it is also true that the two basic assumptions (i.e., fully rational decision–makers and common knowledge of full rationality) upon which normative game theory and its central solution concept (the Nash equilibrium) are built have provoked contradiction and debates which still are ongoing.[38] What it is argued is that such assumptions are too strong ("non realistic" Selten (1994) says) and that, consequently, the predictive power of the theory is questionable. In particular, the quarrel moves along the following lines.

Game theoretic analyses usually take for granted that the notion of 'rational' agent can be assigned a sharp and unambiguous meaning; namely, the meaning associated with rational choice theory, or the (neoclassical) economic approach to social life.[39] A digression on what 'rational' means in this approach appears, hence, necessary.

"A glance at any dictionary will confirm that economists, firmly entrenched in the 'static' viewpoint described above [the viewpoint of substantive rationality], have hijacked this word [rational] and used it to mean something for which the word *consistent* would be more appropriate".[40] Savage's (1954) theory on individual decision–making (in which he synthesised Von Neumann and Morgenstern's expected utility theory with the subjective probability ideas of Ramsey, De Finetti and others) can be considered entirely and exclusively as a *descriptive* theory of *consistent* behaviour. It has nothing to say about *how* decision–makers come to have the tastes and the beliefs ascribed to them; it asserts that if the decisions taken are consistent (in a sense made precise by a list of formal axioms), then the decision–makers act *as if* they were maximisers of expected utility relative to a subjective probability distribution.

The mistake of orthodox game theorists seems to lie in supposing that Savage's passive descriptive theory can be reinterpreted as an active prescriptive theory at negligible cost. In order to come up with precise predictions on rational behaviour, they consider enough to assign prior beliefs to a decision–maker

[37] Some of these critical remarks have been stressed through the exposition, more are discussed by Ledyard (1995).

[38] Objections to these central assumptions as well as to the Nash equilibrium concept are provided, for example, in Binmore (1990); Kreps (1990); Selten (1994); and Hargreaves–Heap and Varoufakis (1995). Most of the discussion which follows rests on these writings.

[39] See Downs (1957); and Becker (1976).

[40] Binmore (1990, p. 152).

and to assume (by following the Aumann–Harsanyi argument) that there is common knowledge of a common prior.[41] Consistency then forces any new information that may transpire to be incorporated into the system by Bayesian updating (i.e., a posterior belief is deduced from the prior belief using Bayes's rule). In this process, the problem of where original beliefs come from has been forgotten. But—it has been argued—without specifying how people come to hold the beliefs ascribed to them, whatever is inferred and deduced from such beliefs is exposed to the danger of being without prescriptive validity.[42]

In addition, there seems to be something distinctly optimistic about the part of the argument which concerns the Harsanyi doctrine: namely, that priors should be taken to be common, to which Aumann adds the rider that it should be common knowledge that the priors are common. This doctrine seems indeed to depend on a powerfully algorithmic and controversial view of reason.[43] Reason on this account is similar to a set of rules of inference which can be used in moving from evidence to expectations. That is why people using reason (because they are using the same algorithms)[44] should come to the same conclusions. However, there is a genuine puzzlement over whether such an algorithmic view of reason can apply to all circumstances. The question is: can any finite set of rules contain rules for their applications to all possible circumstances? The answer given by non–orthodox game theorists is no. According to them, under some sufficiently detailed level of description of an event, there will be a doubt about whether the rule applies to this event. In such a case, an individual will need rules for applying the rule (for applying the rule). Since there is no limit to the details of the description of events, an individual will need rules for applying the rules for applying the rules, and so on to infinity. In other words, every set of rules will require creative interpretation in some circumstances, and therefore, in these cases, it is possible for two individuals who share the same rules to hold divergent beliefs. After all, "we do not seem to expect the same fixtures to be drawn when we complete the football pools; nor do we enjoy the same subjective expectations about the prospects of different horses when some bet on the favourite and other on the outsider".[45] Of course, some of these differences might be due to differences in information, but it is difficult to believe that this accounts for all of them.

There remains, finally, a most important question: *why* a rational decision–

[41] See p. 10 of this work.

[42] See, for example, Binmore (1990).

[43] Cf., Hargreaves–Heap and Varoufakis (1995).

[44] Binmore (1990, p. 145) refers to such algorithms as information–processing tools.

[45] Hargreaves–Heap and Varoufakis (1995, p. 27).

maker should want to be consistent.[46] After all, consistency may lead to inefficient outcomes and players may get high benefits from inconsistencies (as it is the case in public goods settings).

The following statement of Binmore appears to capture quite well all the issues discussed above: "My contention is that the conventional approach misses this aim [to exclude irrelevancies so that attention can be focused on matters of genuine importance], not only by leaving unformalized factors which matter, but also by introducing formal requirements that cannot be defended operationally except in terms of mathematical elegance or simplicity" (Binmore 1990, p. 152).

Therefore, the ambitious claim that game theory will provide a unified foundation for all social sciences[47] seems misplaced to some theorists. They identify a variety of problems with such a claim: some are associated with the assumption of the theory, some come from the inferences which are often drawn from these assumptions, and others more come from the failure (even once the controversial assumptions and the inferences are in place) to generate determinate predictions of what a 'rational' agent would or should do in important social interactions.

Thus, coming back to the question raised at the beginning of this section, it seems reasonable to say that the pioneering experimental research on behaviour in public goods environments (even if not always conducted under controlled or well designed laboratory conditions) mirrors the reported doubts about mainstream game theory and reveals that subjects are more complexly motivated than such a theory allows.

The point now is to understand which are these motivations and what exactly causes cooperation. Modern experimental research is directed toward this aim: its primary purposes are to find out the source of cooperative behaviour and to identify the factors which affect cooperation.

Next chapter will report on what has been discovered so far and, therefore, where the next work might begin.

---

[46] Cf., Binmore (1990).

[47] Cf., Elster (1982); and Aumann and Hart (1992).

Chapter 4

Why do people cooperate if full rationality demands defection? Some hypotheses of cooperative behaviour

4.1 Introduction

In the preceding chapter we looked at the initial laboratory examination of the public goods problem. We found, in contrast with what Nash expected, that in one–shot games and in the initial periods of finitely repeated games people voluntarily provide public goods, and that contributions decline when the decision sequence is repeated. However, repetition seemed to induce decay in certain environments and not in others.[1] The research problem becomes, hence, to discover *when* subjects cooperate and *why* they do so.

Various explanations for individuals' cooperative behaviour have been advanced. The purpose of this chapter is to discuss them and to analyse the way in which these theories have been tested in laboratory.

4.2 A taxonomy of some hypotheses of cooperative behaviour

While the dominant strategy hypothesis does not provide an accurate description of observed behaviour, it has the advantage of being a simple, well–defined argument with clear predictions. By following Bolton et al. (1997) and Bolton

[1] Cf., Isaac et al. (1984).

(1998), I exploit this clarity to classify different theories of cooperative behaviour by minimal modifications to the dominant strategy argument.

4.2.1 Dividing the dominant strategy argument into three parts

The full rationality argument for playing the dominant strategy can be divided into three components:

1. *Motive*: a player prefers having more money to having less.

2. *Cognition*: a player can identify the action that best satisfies Motive.

3. *Choice*: a player takes the action that best satisfies Motive.

In a standard single–shot prisoner's dilemma, for instance, the dominant strategy argument is assembled as follows: each player realises that defection maximises the money he will make, independently on what the other does (motive plus cognition); hence, each player chooses defection (choice).

Most theories of cooperative behaviour depart from full rationality on one of the three components and rely on full rationality (sometimes implicitly) for the other two. In this section, I distinguish among alternative theories by the major component of deviation.

4.2.2 Altruism and equity theories: alternatives to motive

The theories discussed here challenge the game theory's assumption concerning what motivates people. None of them represents a radical departure from the 'more money is preferred to less' assumption. Rather, each supposes an additional motive to interact with self–interest. In all cases, the motive is added as an argument to some form of preference function. Models differ on what is added.

Altruism theories

These kinds of theories posit that individuals care about more than their own self–interest; they also care about the others' welfare.

The usual technique to model altruistic preferences is to add a variable to the classical utility function representing either the consumption or the utility of the others. Collard (1978) distinguishes between commodity–related altruism and utility–related altruism depending on the additional variable used.

Models of altruism have been influential in explaining economic behaviour in many settings, including charitable contributions and volunteer behaviour (e.g.,

Unger (1991); and Smith et al. (1995)), social security and other welfare systems (e.g., Coate (1995)), international bequests and macroeconomic growth (e.g., Rangazas (1991); Hory (1992); Chakrabarti et al. (1993); and Strawcyznski (1994)). Other studies have examined altruism from an evolutionary perspective, either describing evolutionary reasons for altruistic preferences or determining the evolutionary outcomes of societies with heterogenously altruistic individuals (e.g., Bergstrom and Stark (1993); Samuelson (1993); Bergstrom (1995); and Bester and Güth (1998)).

To see how altruistic preferences induce people to contribute an amount greater than the equilibrium, let us consider the symmetric public goods game presented in Chapter 2 (Subsection (2.3.2)). In such a game, a way to insert *pure altruism* into player i's utility function is by writing:

$$V_i = V\big(U(g_1, y), ..., U(g_i, y), ..., U(g_N, y)\big) \tag{4.1}$$

where $y = \sum_{j=1}^{N} g_j$ denotes the total amount of public good provided, and V_i increases in all arguments.

By assuming that the altruist's utility function is linear and that the concern he expresses for the others' welfare is represented by a weight γ, Eq. (4.1) becomes:

$$V(g_i, y) = U(g_i, y) + \gamma \sum_{k \neq i} U(g_k, y) \tag{4.2}$$

where $0 < \gamma \leq 1$.[2] If $\gamma = 0$, then $V(g_i, y) = U(g_i, y)$ and player i shows no concern for the welfare of his partners; this contrasts with altruistic preferences. If $\gamma = 1$, the altruist treats the others as himself.

Assigning the functional form (2.4) to all $U(g_j, y)$ $(j = 1, ..., N)$ in (4.2), we have:

$$V_i = (m - g_i) + vy + \gamma \sum_{k \neq i} (m - g_k) + \gamma \sum_{k \neq i} vy,$$

where $(m - g_k)$ are the benefits that i's kth fellow member derives from his private consumption. The term $\sum_{k \neq i}(m - g_k)$ is, therefore, a constant from i's point of view and we can indicate it with Pr_k. Thus, the latter equation can

[2] Edgeworth (1881, p. 53) employed a similar formulation of altruism. Restricting attention to two people (i and j), he described i's altruistic preferences by:

$$V_i = \gamma U_i(g_i, g_j) + (1 - \gamma) U_j(g_i, g_j),$$

and called the value $(1 - \gamma)/\gamma$ the 'coefficient of effective sympathy'.

be written as:

$$V_i = (m - g_i) + vy + \gamma(N-1)vy + \gamma Pr_k,$$

or, equivalently

$$V_i = (m - g_i) + \gamma Nvy + (1-\gamma)vy + \gamma Pr_k. \tag{4.3}$$

Eq. (4.3) implies that, whenever the inequalities stated in (2.5) hold (i.e., when the parameter v satisfies the constraints $v < 1$ and $Nv > 1$), if the weight is such that $0 < \gamma < 1$, then the optimal contributions for an altruist will be greater than zero and less than m. They will coincide with m, when $\gamma = 1$.

Public goods models that incorporate the specification of altruism given in (4.2) have a neutrality property. Specifically, the models imply that government contribution to the public good, funded by lamp–sum taxes, crowds out private contribution with a one–to–one correspondence. The argument can be stated briefly in terms of function (4.2) by assuming $N = 2$. Imagine that i is the only contributor. Since—from (4.2)—i's preferences are completely determined by the total amount of public good provided y, i should be indifferent to whether he contributes g_i voluntarily or involuntarily through a tax transfer. Hence, if a law forces him to contribute $\tilde{g}_i < g_i$, i's voluntary contributions should fall by exactly $\tilde{g}_i$. It has been shown that this one–to–one crowding out (or neutrality property) holds also when altruistic preferences are modelled differently from (4.2),[3] and even for some distortionary taxes.[4]

Field studies of charitable giving, however, report that actual crowding out is quite small. One explanation for this inconsistency, advanced by Andreoni (1989), is known as *impure altruism hypothesis.* Specifically, Andreoni's argument is that people receive some utility—a 'warm–glow'—from their voluntary gift *per se.* By introducing a warm–glow parameter as additional argument into altruistic utility functions, Andreoni showed that the resulting model is consistent with the empirical results.

Equity (or inequality aversion) theories

The ERC model by Bolton and Ockenfels (2000) posits that people are motivated by their own pecuniary payoff as well as by their own relative payoff, "a measure of how a person's pecuniary payoff compares to that of others". ERC stands for **e**quity, **r**eciprocity and **c**ompetition: the authors show, indeed, that

[3] See Sugden (1982, p. 346) for a proof.

[4] Cf., Bernheim (1986).

their model can account for phenomena reported in bargaining (equity), market (competition), and dilemma (reciprocity) games.

Bolton and Ockenfels characterise preferences in terms of what they call *motivation function*, which "may be thought of as a special class of expected utility functions". The term 'motivation function' is preferred because it makes clear that the emphasis is on the objectives that motivate behaviour during the experiment, and (although the weights that individuals put on these objectives may not be immutable) the important is that the trade–off between pecuniary and relative payoffs remains stable for the duration of the experiment. In an N-player game with monetary and nonnegative payoffs y_i, each player i is assumed to act in order to maximise the expected value of his motivation function:

$$\nu_i = \nu_i(y_i, \sigma_i),$$

where σ_i is i's relative share of the payoff:

$$\sigma_i = \sigma_i(y_i, c, N) = \begin{cases} y_i/c & \text{if } c > 0 \\ 1/N & \text{if } c = 0, \end{cases}$$

and $c = \sum_{j=1}^{N} y_j$ is the total pecuniary payout distributed among all players. We have, hence, $y_i \equiv c\sigma_i(y_i, c, N)$.

How a lab subject trades off pecuniary and relative payoffs is clearly private information. On the other hand, testing the model requires a reliable (preferably observable) measure of the underlying trade–offs. With respect to this point, the authors found that much of what we need to know is captured by the thresholds at which behaviour deviates from the "more money is preferred to less" assumption. For all $c > 0$, each player is supposed to have two thresholds, $r_i(c)$ and $s_i(c)$, defined as follows:

$$r_i(c) = \underset{\sigma_i}{\text{argmax}} \quad \nu_i(c\sigma_i, \sigma_i);$$

$s_i(c)$ is implicitly defined by

$$\nu_i(cs_i, s_i) = \nu_i(0, 1/N), \qquad s_i \leq 1/N.$$

Bolton and Ockenfels demonstrated that knowing the distributions of these thresholds is sufficient to characterise many phenomena. To see, for instance, how their model justifies cooperation in dilemma games, consider the prisoner's dilemma payoff matrix displayed in Table 4.1.

Table 4.1: Prisoner's dilemma payoff matrix.

	Player 2	
Player 1	Cooperate	Defect
Cooperate	$2m,\ 2m$	$m,\ 1+m$
Defect	$1+m,\ m$	$1,1$

m = marginal per capita return $\in (0.5, 1)$.

Suppose that player i $(i = 1, 2)$ can be described by the following additively separable motivation function:

$$\nu_i(c\sigma_i, \sigma_i) = a_i c\sigma_i - \frac{b_i}{2}\left(\sigma_i - \frac{1}{2}\right)^2. \tag{4.4}$$

The component $a_i c\sigma_i$ is an expression of standard preferences for the pecuniary payoff. The other component of (4.4) delineates the influence of the comparative effect.[5] Therefore, a/b (the ratio of weights attributed to the pecuniary and relative components of the motivation function) fully characterises a player's type. Strict relativism is represented by $a/b = 0$; strict narrow self–interest is the limiting case $a/b \to \infty$. Let us denote by $F(a/b)$ the population distributions of type. It will be optimal for a subject with type a/b to cooperate rather than defect if:

$$\frac{a}{b} < \frac{p - \frac{1}{2}}{4(1-m)(1+2m)^2} = g(m, p)$$

where p is the probability that the opponent cooperates, and m is the MPCR.

Thus, in the ERC model, cooperation is influenced by the extent to which subjects are motivated by relative payoffs, the magnitude of the MPCR, and the proportion of cooperating subjects in the population. There is always an ERC equilibrium in which no one cooperates, but—depending on the shape of $F(a/b)$—there can also be ERC equilibria in which a proportion of subjects cooperate, while others defect.[6]

Fehr and Schmidt (1999) formulate a theory of *inequality aversion* which makes predictions similar to those of ERC. Even the Fehr–Schmidt model is

[5] Basically, the further the allocation moves from i receiving an equal share, the higher the loss from the comparative effect.

[6] An ERC equilibrium is defined as a perfect Bayesian equilibrium solved with respect to player motivation functions where each player's r and s are private information but the density functions f^r and f^s are common knowledge.

based on the assumption that (to some extent) people dislike inequality in payoffs, and that they dislike inequality more if it is to their disadvantage than if it is to their advantage. Like in the ERC model, such an assumption is captured by incorporating an equity term along with an own–earnings term into an individual's utility function. Applied to public goods situations, the model predicts that, as long as inequality–averse players believe that other players are contributing, they are willing to contribute too.

There are two differences between the ERC model and the Fehr–Schmidt model. One is that, in the latter, inequality between self and all others matters instead of inequality between self and the average earnings of others. The other difference is that most of the results in the Fehr–Schmidt model are derived in a complete information context, whereas the ERC is an incomplete information model. Both models are, however, based on preference assumptions which transform the dilemma game (with its unique inefficient equilibrium) into a coordination game (with multiple equilibria), where players have to form beliefs about the others' choices in order to choose one of the equilibria.

4.2.3 Reciprocity and commitment theories: alternatives to choice

Theories which reformulate the choice rule of orthodox game theory are traditionally considered as bounded rationality theories. They move away from the maximisation assumption and assert that people forgo their own self–interest because of a sense of obligation to conform to particular rules. These models differ from one another with respect to how far they move away from formal optimisation as well as with respect to the alternative decision rule considered.

Reciprocity theories

Theories of reciprocity assume that individuals reciprocate or match the others' contributions.

Sugden (1984) proposed a model based on the idea that a subject who benefits from the public good feels a moral obligation towards those who contribute which leads him to contribute as well. Specifically, in his model, a person maximises his own utility subject to an external moral constraint: the 'principle of reciprocity'. Sugden writes player i's utility function as:

$$U_i = U_i(q_i, z)$$

where q_i is i's contribution to the public good and z is the quantity of public

good provided. U_i increases in z and decreases in q_i. The constraint can be stated as follows. Let q_i^G be the contribution that maximises U_i under the assumption that everyone contributes the same amount. Then, i has the obligation to contribute at least q_i^G if everyone else does so, or at least q_j if some person j in his group contributes an amount q_j less than q_i^G. Formally, i is meeting the reciprocity principle (i.e., his obligation towards the group) if and only if *either* $q_i \geq q_i^G$ *or*, for some other person j, $q_i \geq q_j$. An equilibrium is a vector of contributions in which each player contributes the smallest amount that meets his obligation.

One implication of the model is that an increase in one person's contribution tends to induce others to increase their own contributions.[7]

Commitment theories: Kant and morality

A further suggestion for explaining cooperation connects rationality with morality and Kant provides a ready reference. His practical reason demands that we should undertake those actions which when generalised yield the best outcomes. Or, to say the same thing slightly differently, according to Kant, an agent should undertake only an action that can be generalised to good effect by all agents. It does not matter whether others perform the same calculation and actually undertake the same action as you. The morality is deontological and it is rational for an agent to be guided by a categorical imperative.

As an example of how the categorical imperative might be applied and how it differs from full rationality, consider a person wondering whether to pay his taxes. Non–payment could be rational (in a standard sense) in so far as the person is interested only in his welfare and the chances of being fined for non–payment are low. However, such a behaviour would not pass the test of the categorical imperative. If the person were (hypothetically) to consider not paying while at the same time accepting the premise that others are similarly rational, then he would be committed to the predictable result that society would break down and collapse without the necessary funding. For Kant the rational person should not allow his reason to be a slave to the passions (which might lead to non–payment); instead his rationality, and the fact that he shares it, should lead him to the categorical imperative and hence to the payment of the taxes.

This is perhaps the most radical departure from the conventional understanding of what is required by rationality because, while accepting the payoffs, it suggests that agents should not act in a traditional way upon them. The no-

[7] See Result 4 of Sugden (1984, p. 780) for a proof.

tion of rationality is no longer understood in the means–end structure (as the choice of the action most likely to satisfy given motives). Instead, rationality is conceived as an expression of what it is possible; it has become a motive in its own right.

I shall refer to the class of theories based on Kantian reasoning as *commitment theories*.[8] In the spirit of the categorical imperative, these theories assume that, even if each individual is motivated by purely egoistic satisfaction derived from the goods that he possesses, there is an implicit social contract such that each performs duties for the others in a way calculated to increase the welfare of all.

Laffont (1975) analysed the case where each individual believes that the others will act as he does, and then he maximises his utility given that belief. Laffont showed that, under these beliefs, individuals voluntarily contribute and social welfare increases.[9]

Similarly, Harsanyi (1980) described the principle of 'rational commitment' in which an individual makes whatever contribution he would wish others to make, irrespective of whether they actually make this contribution.

4.2.4 Reputation building, errors and learning theories: alternatives to cognition

Models which modify the cognition component of the full rationality argument assume that people are not able to recognise the action that maximises their utility *either* because their information about the others is incomplete *or* because they do not grasp the incentives of the game. In the case of incomplete information, early contributions may be consistent with rational behaviour. In the case of not understanding of the game, people's behaviour can be characterised in terms of errors or in terms of a learning process in which the rule that governs the choice changes with experience.

Reputation building and the strategies hypothesis

It is the claim of both Kreps et al. (1982) and McKelvey and Palfrey (1992) that a form of cooperation is rationally possible when the game is repeated.[10] Both models alter the classical theory by a single precept: instead of supposing

[8]This expression is used by Croson (1998b). Collard (1978; 1983) called these theories 'Kantian', and Sugden (1984) referred to the principle underlying them as the 'principle of unconditional commitment'.

[9]He also discussed the social benefits of a government convincing the population that this belief is true.

[10]See also Kreps and Wilson (1982); and Milgrom and Roberts (1982).

that it is common knowledge that all players are rational, they suppose that the information about the types of players is incomplete, and that individuals attach a positive probability to the likelihood that the others are not fully rational (i.e., they use dominated strategies).

Kreps et al. (1982) described an equilibrium in the finitely repeated prisoner's dilemma in which two rational players both believe that there is a small probability, δ, that the other is 'irrational' and, as a consequence, they *rationally* build reputation. Kreps et al. gave two examples of irrationality. First, the opponent might be playing a tit–for–tat strategy, which begins by cooperating and then plays whatever the other played on the last round. Second, players could believe that the opponent gets extra utility from mutual cooperation, such that cooperation is the best response to cooperation. In each case, a sufficiently high δ can lead each player to adopt a strategy of the sort "cooperate until round T, or until the opponent defects, and defect thereafter". Higher values of δ will tend to increase the amount of cooperation.

A strict interpretation of Kreps et al.'s theory is that no irrational or altruistic types need to exist, but only that there are sufficient beliefs that such types exist. However, since in the known last period defecting is always optimal, in anticipation of this subjects may start 'bailing–out' and, as it becomes increasingly clear that the population is all rational, cooperation may become increasingly difficult to maintain.

Andreoni (1988) first applied Kreps et al.'s theory to public goods experiments and referred to it as the *strategies hypothesis*. According to this hypothesis if a (perfectly rational and selfish) player is not sure whether the other players are fully rational, then, in early repetitions of the game, he may find convenient not to educate them to play the dominant strategy; rather, he may have an interest in building up a reputation as a cooperative type himself. This would imply a relatively high contribution level in the early periods, which decreases either as the player has no longer doubts about the others' rationality (because, for instance, he observes them playing the dominant strategy), or in the (known) last period of interaction, when it does not make any more sense to build up reputation (since there is not future, contributing to the public good cannot induce any further contributions by the others).

The strategies hypothesis is quite vague about the number of periods needed by a player to dissolve all his doubts about the others' rationality. Andreoni does not make any specific assumption with respect to this point and, at the best of my knowledge, no experimental study has examined this issue. However, since my analysis of the strategic type and the way in which the strategies hypothesis

is treated in this work require to know when exactly a player stops doubting about the rationality of his partners, throughout the thesis, I shall assume that the latter happens whenever an agent observes his partners to play the dominant strategy for two consecutive periods. Such an assumption is my own and the reader may well find that he (or she) disagrees.

Theories of errors: confusion and simple learning

A further explanation for contributing behaviour can be found in the *inexperience* of the subjects. The lack of familiarity with the game may lead people to contribute out of confusion or error.

Two different mistakes hypotheses have been proposed. One asserts that mistakes are *random* and derives from Ledyard's (1995) proposition that "subjects make mistakes, do not care, are bored, and choose their allocation randomly" (Ledyard 1995, p. 170). The other asserts that mistakes are *systematic*, in the sense that the players confuse the game they are in with one that is more common to their everyday experience. Failing to recognise the dominant strategy, they contribute because this is the social norm that they associate with the 'game of life' (i.e., with the everyday game). This hypothesis—called *misidentification* by Bolton et al. (1997)—is a specific interpretation of the argument advanced by Gale et al. (1995) in the context of ultimatum games. Notice that misidentification differs from random errors in two ways. First, it postulates a systematic pattern of contribution that random error does not. Second, it may be strategy sensitive in the sense that it may vary depending on the information about the others' contributions; Ledyard's statement of the mistakes hypothesis, on the contrary, is strategy insensitive.

Palfrey and Prisbey (1992) noticed that when the Nash equilibrium is a corner solution, the only way in which a subject can err is to contribute too much to the public good. Hence, contributions due to 'background noise' have been misinterpreted as purposeful contributions and this has led some authors to overstate the importance of altruism or strategic reputation–building play.[11] Palfrey and Prisbey turned their observation into a new hypothesis; namely, that *confusion* among the players might explain contributing behaviour.

Andreoni (1988) proposed the *simple learning hypothesis* as an explanation of the decay phenomenon observed in public goods experiments. According to this hypothesis, subjects might not immediately understand the incentives of the game, but need time in order to realise that not contributing is the dominant strategy. Andreoni (1995a) argued that, only if confusion is the

[11] See also Ledyard (1995, p. 147).

principal explanation for cooperation, the emphasis on learning in the literature is justified.

Reinforcement learning models

If subjects are trying to learn (by some suitable groping process) what the appropriate strategy is, then a learning algorithm would provide the right model for describing the process.[12] If everyone learns, one should observe that contributions converge to the non–cooperative equilibrium after enough periods. This seems to happen after 10 iterations in small groups. We do not know, instead, how long it would take in large groups.

Roth and Erev (1995) described a simple reinforcement learning algorithm predicting the dynamic path of play as people learn about the game and about the others' behaviour. The reinforcement was introduced to increase the frequency of those strategies which had more success in the past periods.

The basic mechanics of the Roth–Erev model were as follows. Each player i begins the first round of the play, $t = 1$, with an initial propensity to play his kth pure strategy given by some number $q_{ik}(1)$. Repeated play modifies these propensities through a process of adaptation. If player i plays his kth strategy in round t and receives a payoff x, then the propensity to play k in round $t + 1$ is updated to $q_{ik}(t + 1) = q_{ik}(t) + x$; that is, the reinforcement which a player receives after playing a strategy k and receiving a non–negative payoff x is set equal to the payoff itself.[13] The probability that k gets played in round t is $q_{ik}(t)/\sum q_{ij}(t)$, where the sum is taken over all pure strategies j of i. The predictions of the model were derived from computer simulations of the reinforcement process. The intermediate run of the average simulation path was then compared to the actual path observed in the experiment.

The Roth and Erev model incorporates some of the robust properties of learning typical of the psychology literature. Specifically, the *Law of Effect* which states that the choices that have led to good outcomes in the past are more likely to be repeated in the future, and the *Power Law of Practice* which states that learning curves tend to be steep initially and then flatter.

Roth and Erev took also into account the public good game known as 'best shot'. In this two–person game, the first mover chooses a contribution of tokens. After viewing the first mover's choice, the second mover makes a contribution. The maximum of the two determines the total cash to be divided: the larger the

[12] These types of models have been proposed by Boylan (1990); Crawford and Haller (1990); Kalai and Lehrer (1990); Miller and Andreoni (1991); and Roth and Erev (1995).

[13] This way to model reinforcement was due to the fact that, in the three games in which Roth and Erev tested their model, all potential payoffs were non–negative.

maximum, the larger the pie. The largest payoff, however, goes to the player who contributes the least. Perfect equilibrium has the second mover doing all contributing and receiving, therefore, much less than the first mover. Along the equilibrium path, the best shot second mover has the opportunity to end the game with both players receiving nothing.

Studies by Harrison and Hirshleifer (1989) and Prasniskar and Roth (1992) found that, after a few iterations, best shot play approaches 100% perfect equilibrium. This outcome was anticipated by the Roth–Erev model.

Duffy and Feltovich (*forthcoming*) provided further evidence for reinforcement learning in best shot games. Miller and Andreoni (1991) demonstrated that *replicator dynamics* can explain some of the stylised fact in public goods experiments.[14]

In their model, Roth and Erev did not attempt to predict initial playing propensities which were taken as given.[15] Nevertheless, the way in which these initial propensities are interpreted is important if one wants to compare reinforcement learning with altruism, reciprocity and commitment theories. The crucial difference between reinforcement learning and these other theories seems to lie on the fact that the former assumes that all concerns for altruism, reciprocity or Kantian reasoning will dissipate with enough reinforcement of the right sort.

4.3 How well do these theories perform in laboratory settings?

The present section reports on a series of experiments which have been run in order to verify the descriptive power of the theories of cooperative behaviour described above. I shall start the discussion by taking up results dealing with repetition and the related issues of learning and experience. Then, I will focus on experiments expressly designed to separate confusion (or errors) from cooperative or altruistic theories. Finally, since it will be shown that linear altruism models are not sufficient to account for all cooperation observed in public goods settings, I will examine the role and the weight (in comparison also with altru-

[14] A replicator dynamic is a simple and canonical dynamical representation that biologists use to formalise the genetic mechanism of natural selection. In this dynamic, the fraction of the population playing strategy k is determined by how well k fares relative to the population average fitness. Note that this argument rewards strategies that do well against the population, not those that are the best.

[15] Roth and Erev first analysed the model using initial conditions drawn randomly from a uniform distribution, and then compared the model to experimental data using initial conditions fitted from the first round of play.

ism hypotheses) of framing, reciprocation and equity in explaining cooperation.

4.3.1 Repetition, learning and strategies

For those environments where contribution rates diminish after some number of iterations, the relevant point is to find out whether such a decay is due to learning or strategic behaviour. In analysing the experimental literature which tried to address the question of the relative importance of learning and strategies, there are a number of different dimensions to which we need to pay attention: 1) the Partners/Strangers design; 2) the restart effect; 3) the subjects' end behaviour.

The Partners/Strangers design

The truly innovative contribution of Andreoni's 1988 paper was the introduction of the Partners/Strangers design. In an Isaac and Walker environment with $p = 1$, $a/N = 0.5$, $N = 5$, and $m = 50$,[16] Andreoni compared two different treatments. In one, the *Partners* treatment, group members remained fixed throughout all the session, which consisted of a sequence of 10 decision–rounds. In the other, the *Strangers* treatment, group members were randomly re–assigned to new groups of 5 after each repetition.

Andreoni's idea was to separate strategic play by Partners from non–strategic play by Strangers. In the Strangers condition, in fact, one should see only learning: any increase in free–riding that is observed in such a condition cannot be attributed to strategic play but to the fact the subjects learn from their own experience that free–riding is the dominant strategy. On the other hand, any differential increase in free–riding that is observed between the two treatments can be attributed to strategic play, but not to learning. Suppose a subject is initially investing some positive amount in the public good, but learns in round t that free–riding is the single–shot dominant strategy. If he is a Partner—playing strategically—he may continue to contribute to the public good; instead, if he is a Stranger, he has no incentive to continue cooperation (every game for him is, after all, an end–game). Therefore, if one believes in the strategies hypothesis, one should expect that giving by Partners will be greater than giving by Strangers, especially early in the game.

Results of Andreoni's experiment are reported in Table 4.2.

As can be seen, contrary to expectations and to the strategies hypothesis, subjects who could not play strategically actually provided more of the public

[16] See p. 43 of this work.

Table 4.2: Andreoni's (1988) experiment: Average investment in public good per subject.

	Round										
	1	2	3	4	5	6	7	8	9	10	All
Partners	24.1	22.9	21.5	18.8	18.4	16.8	12.8	11.2	13.7	5.8	16.6
Strangers	25.4	26.6	24.3	22.2	23.1	21.9	17.8	19.7	14.0	12.2	20.7
Difference	-1.3	-3.7	-2.8	-3.4	-4.7	-5.1	-5.0	-8.5	-0.3	-6.4	-4.1

good than subjects who could. In all ten rounds of the experiment, in fact, Strangers contributed more than Partners.

I have, however, some doubts about Andreoni's results. In particular, the Strangers treatment does not seem to be correctly designed: it goes to allow for strategic play. Indeed, since the experiment involved 20 subjects who, divided into 4 groups of 5, played the game ten times and since this structure was common knowledge, Strangers knew they would have met one another more than once. Such a knowledge may have, of course, influenced their behaviour inducing them to play strategically. Hence, any result concerning the learning hypothesis which is drawn in such a design does not appear extremely convincing. As a consequence, also the possibility of using the comparison between the Partners and Strangers treatments to throw light on strategic play weakens.

Andreoni's finding (that Partners contribute less than Strangers) has, in fact, shown itself not to be robust to replication.

For instance, in an experiment directed to examine the Partners/Strangers treatment, Weimann (1994) found no statistically significant difference between Partners and Strangers contributions. Specifically, only in 4 out of 30 cases significant differences were observed; in 3 out of these 4 cases, Partners made significantly higher contributions. This is at odds with Andreoni's results.

However, the methodology used by Weimann for particular treatments was substantially different from that used by Andreoni. In Weimann's experiment, the Partners treatment was run in laboratory, while the Strangers treatment was run by phone and involved no personal contact with either the experimenter or the other subjects. In addition, half of Weimann's subjects received information about the individual contributions of each member of their group, rather than information only about the total contribution of the group, as in Andreoni. While average contributions are indistinguishable in these two treatments, related work by Sell and Wilson (1991) shows that the variance of contributions differs substantially. Thus pooling observations from these treatments

as Weimann did is misleading.

One more study that replicated Andreoni's experiment was carried out by Croson (1996). Her design differed from that of Andreoni in only two aspects: the group size (4 rather than 5) and the subjects' initial endowment (25 rather than 50). Nevertheless, her results were in sharp contrast with those of Andreoni. In Croson's experiment, Partners always contributed more than Strangers, with the difference between the two narrowing as the end of the game approached. Hence, all her observations are compatible with the strategies hypothesis.

Testing the robustness of Andreoni's findings was also part of the purpose of Burlando and Hey's (1997) experiment, which was run in two different geographical places: in England (York) and in Italy (Turin). Burlando and Hey cast their experiment in terms of a public 'bad'. The results they obtained were decidedly mixed: the Partners/Strangers distinction by itself was not significant. It appeared instead relevant when the data were split by nationality: for the English subjects, Partners free–rode rather more than Strangers (just like Andreoni discovered) although the difference in contributions was not significant; for the Italian subjects, on the contrary, Partners free–rode less than Strangers. These findings give the impression that national differences (which, as the authors suggest, presumably reflect cultural, sociological and psychological differences between subject groups) have a strong effect on individual behaviour.[17]

The difference in results between the considered studies highlights the importance of replications for economics experiments. When we are searching for robust behaviour, only by replicating experiments we can safely drawn conclusions about our investigations.

The restart effect

A second innovative feature of Andreoni's 1988 paper was that of introducing—and studying the effects of—a *restart*. The basic experiment was performed twice. After the announced end of the game, subjects were unexpectedly told that they would restart a new set of 10 rounds. Partners stayed in the same group, while Strangers continued to be randomly re–assigned. However, Andreoni suspended the restart game after only three additional rounds.

A short break between two subsessions allows for what is called in psychology *cognitive dissonance*, a moment in which players stop the continuity of

[17] Ledyard (1995) provides an excellent summary of the factors known to influence the level of public goods contributions.

decisions and/or actions and rethink about what to do next in the light also of the recent experience.[18]

Such a procedure was introduced in order to isolate the learning hypothesis. If learning is primarily responsible for decay, then the subjects—both Partners and Strangers—should be unaffected by the restart and their eventual trend towards the free–riding equilibrium should continue.

At odds with the simple learning hypothesis, Andreoni observed a strong restart effect in the Partners treatment,[19] and a similar but weaker effect in the Strangers treatment.

After reporting his results, Andreoni concluded that "neither strategies nor learning can be supported as explanations of decay in public goods experiments" (Andreoni 1988, p. 300), and that, in order to clarify such a phenomenon, we need to consider a richer set of behavioural motivations such as altruism, social norms, or regret.

One problem in Andreoni's experimental design is that his restart was not announced at the beginning of the experiment and the restart announcement (i.e., that the experiment would be replicated) was not carried through to completion. As pointed out in the previous chapter (and as Andreoni himself argues),[20] such deceptive practices are not recommended in economics experiments.

Accordingly, both Croson (1996) and Burlando and Hey (1997) built in a restart into their procedures from the beginning. In their experiments, subjects were explicitly told that they would play two subsessions with a short break in the middle.

In Croson's study, the restart effect was significant in the Partners condition,[21] and insignificant but present in the Strangers condition.[22] Both these results are consistent with those of Andreoni and contrary to the simple learning hypothesis. Rather, according to Croson, they seem to provide evidence in favour of strategies.

While Croson's experiment was a replication of Andreoni (1988) in which all ten restart rounds were reported, Burlando and Hey slightly altered Andreoni's original design. Particularly, they allowed for a Partners (Strangers)

---

[18] Akerloff and Dickens (1982) speak about the relevance of cognitive dissonance in economics. The classical reference for the phenomenon is Festinger (1957).

[19] Partners' contributions increased sharply from the last round of the original game to the first round of the restart game.

[20] Cf., Andreoni (1988, p. 295).

[21] Specifically, the average contribution in period 10 of the original game was 4,54 tokens, while that in period 1 of the restart game was 11.54.

[22] Here, the average contribution was 2.48 in round 10 of the original game, and 6.21 in round 1 of the restart game.

subsession following a Strangers (Partners) subsession and, if having two Partners subsessions (before and after the break), for a changing of the Partners in the second one. Such a design was adopted in order to study various effects, including the so–called *sequencing effect*[23] and the relative strength of the restart phenomenon when a Strangers session follows or when a Partners session follows.

Table 4.3 shows Burlando and Hey's experimental results in all various treatments.

Table 4.3: Burlando and Hey's (1997) experiment: Differences in percentages dumped before and after the restart.

	Total	York	Turin
Partners then Strangers	−4.5	−12.8	+3.8
Strangers then Partners	−3.9	−5.3	−2.5
Partners then Partners [a]	−16.3	−16.7	−15.9
Partners then Partners [b]	−0.6	−5.6	+4.4
Strangers then Strangers	−1.2	−2.4	0.0

[a] Same partners.

[b] Different partners.

A look at the data reveals that the greatest reduction in free–riding occurred when the two subsessions were both Partners sessions and, in particular, when the Partners in the second subsession were the same as in the first one; this abets the strategies hypothesis. In contrast, the smallest fall in free–riding occurred when again both subsessions were Partners but the members of the group changed from the first subsession to the second. The second smallest restart effect was when both subsessions were Strangers sessions; this is clearly at odds with the simple learning hypothesis.

Burlando and Hey's findings confirm, therefore, those previously found by both Andreoni and Croson and cast further doubts on the relevance of simple learning as an explanation for the observed decay in cooperation. Rather, their results appear to suggest that "subjects well understand both the potential benefits of socially optimal behaviour and the importance of their fellow–partners understanding it also... With different partners, or with a new set of strangers, subjects could not be sure that other subjects appreciated the potential benefits from mutual cooperation, or were less inclined to bet on their group mates'

[23] Namely, whether subjects' behaviour in either Strangers or Partners sessions is influenced by prior experience in a different session.

willingness to cooperate" (Burlando and Hey 1997, p. 57). This leads Burlando and Hey to conclude that: "... subjects are less concerned with strategic play against their partners and more concerned in understanding how their fellow subjects perceive the game" (Burlando and Hey 1997, *ibidem*).

The final–round effect

According to theory, any attempt of socially desirable behaviour should collapse completely in the final round of any session: in the Partners session, the scope for strategic behaviour disappears; in the Strangers session, all subjects should have learnt that zero contribution is the dominant Nash equilibrium.

Andreoni found that contribution rates were least in the final round (the 10th) both for Partners and for Strangers, but still above the zero investment level. Furthermore, in his experiment, the decay of average contributions in the last round was stronger in the case of Partners than in the case of Strangers.

Weimann corroborated Andreoni's results. In his experiment, there was a clear final–round effect not leading to a (pure) free–riding outcome, and Partners showed a significantly stronger decay in their last round's contributions than Strangers.

As far as Croson's experiment is concerned, she found that in the final period of both the original and the restart games average contributions reached their minimum and that the final–round effect was stronger in the Partners groups than in the Strangers groups.

In Burlando and Hey's experiment, if one considers as final period the end of the restart game (which is, indeed, the true end of the game since their subjects knew they would play again after the break), one has a confirmation of the increase in free–riding in the last round in comparison with the previous round. As to the Partners–Strangers distinction, instead, Burlando and Hey observed more cooperation from Strangers than Partners only in the last round of the first subsession (the original game), while the opposite was true for the second subsession (the restart game).

Besides a confirmation of the importance of repetition on behaviour, the main points emerging from the four experiments described in this section are that cooperation prevails in public goods games against the Nash prediction, and that the simple learning hypothesis cannot be supported as an explanation of the decay phenomenon observed in these kinds of games. To the contrary, the reported observations appear to be compatible with different possible forms of strategic and non–standard behaviour,[24] though Andreoni's results and some

[24] If, by following Harsanyi, we define the behaviour of fully rational individuals as 'standard',

findings of Burlando and Hey seem to contradict the reputation–building hypothesis.

With respect to the various theories of cooperative behaviour, although these four experimental studies find space and scope for them, they were not designed to shed light on any particular hypothesis or to provide settings for comparing them.

4.3.2 The noise versus altruism hypotheses

This section concentrates on four experiments expressly designed to separate confusion from alternative cooperative theories.

The possibility of confusion being the explanation for the findings reported in Subsection (4.3.1) was first proposed by Palfrey and Prisbey (1992). They asserted that the presence of altruism does little to explain the counter–intuitive results in Andreoni (1988) since in that experiment there was a lot of statistical variation across trials, which suggests that the data were noisy. Because of the usual experimental designs, 'noise' (in the sense of statistical deviation from the theoretical prediction) could only manifest itself as overcontribution. As a consequence, the importance of systematic findings such as altruism and strategic play has been overstated. Instead, according to Palfrey and Prisbey, most of the observed anomalies can be accounted for as background noise.

To prove their point, they run an experiment in which each agent's marginal rate of substitution (hereafter, mrs) between the private and the public good was varied throughout the different periods. The results which Palfrey and Prisbey obtained in this early study appear to be indecisive, but they continued on this line of enquiry with clearer experiments in later papers.[25]

To address the questions posed by misunderstanding and confusion, Andreoni (1995a) devised another experiment which represents the first systematic attempt to separate the hypothesis that cooperation in public goods experiments is due to kindness, altruism, or warm–glow from the hypothesis that cooperation is simply the result of errors or confusion.

In this paper, Andreoni discarded reputation–building as an explanation for the lack of the dominant strategy in the laboratory (since subjects were found cooperating also in Strangers sessions) and concentrated only on two hypotheses: 1) the *kindness hypothesis*, according to which the free–riding hypothesis,

---

we can correctly argue that all models analysed in Section (4.2)—which are outside of the classical, full rationality approach—provide examples of models of non–standard behaviour.

[25] Palfrey and Prisbey's 1996 and 1997 papers will be discussed in depth later in this section.

in its pure form, is incomplete and subjects have tastes for cooperation;[26] and 2) the *confusion hypothesis*, according to which subjects have somehow not grasped the true incentives, due either to the experimenters' failure to adequately convey them or simply to their incapability of deducing the dominant strategy.

To distinguish between these two hypotheses, Andreoni compared a *Regular condition* (i.e., a standard public goods experiment) with one in which social, cultural and strategic incentives to cooperate were subtracted off, thus leaving confusion as the most reasonable explanation for cooperation. His design included, in fact, a second condition, the so–called *Rank condition*, in which subjects were paid according to their experimental earnings rank in comparison to the other subjects in their group rather than according the classical voluntary contribution mechanism. Specifically, the subject with the highest experimental earnings got the highest monetary payments, with payments decreasing with rank; if there were ties, the payoffs were split among those who tied.[27] Note that such a payment scheme converts the standard positive–sum public goods game into a zero–sum game.

The important feature of the Rank condition is that, while preserving the dominant strategy equilibrium of the Regular condition (the way to get the highest rank in the group is to be the biggest free–rider, i.e, to contribute zero), its reward scheme offers no incentives for cooperation (if three subjects cooperate they can all raise their own experimental earnings, but the experimental earnings of the other subjects will increase by even more).

Since in the Rank condition subjects had information about their rank and were paid according to rank (whilst in the Regular condition subjects did not know their rank and got paid their experimental earnings), and since the rank information—apart from the payment by rank—could alter behaviour, Andreoni introduced a third condition, the *RegRank*, for control. In this third condition, subjects had the same information on their rank as the Rank subjects, but they were paid according to their experimental earnings like the Regular subjects.

Hence, Rank and RegRank conditions were identical except for the method of payment; as a consequence, the difference in cooperation between these two conditions provides an estimate of the number of subjects who understand

[26] While there are various specific alternatives which may capture this (such as pure altruism or warm–glow giving), Andreoni was not interested in discussing or investigating on them. He wanted only to distinguish the relevance of such an attitude, besides the specific reasons for it or the forms it may take.

[27] Paying subjects according to their rank was also done by Bolton (1991).

the incentives but cooperate out of kindness. On the other hand, since the difference in cooperation between Regular and RegRank conditions was only due to information on rank, it can be attributable to either kindness or confusion.

The top panel of Table 4.4 lists the percentage of free–riders in any one round of Andreoni's experiment.

Table 4.4: Andreoni's (1995a) experiment: Percentage of subjects contributing zero to the public good per round.

	Round									
Condition	1	2	3	4	5	6	7	8	9	10
Regular	20	12.5	17.5	25	25	30	30	37.5	35	45
RegRank	10	22.5	27.5	40	35	45	50	67.5	70	65
Rank	35	52.5	65	72.5	80	85	85	85	92.5	92.5
Kindness:										
Rank-RegRank	25	30	37.5	32.5	45	40	35	17.5	22.5	27.5
As % of 100-Regular	31.3	34.3	45.5	43.3	60.0	57.1	50.0	28.0	34.6	50.0
Confusion:										
100-Rank	65	47.5	35	27.5	20	15	15	15	7.5	7.5
As % of 100-Regular	81.3	54.3	42.4	36.7	26.7	21.4	21.4	24.0	11.5	13.6
Either:										
RegRank-Regular	-10	10	10	15	10	15	20	30	35	20
As % of 100-Regular	-13.0	11.4	12.1	20.0	13.3	21.4	28.6	48.0	53.8	36.4

As can be seen, Rank subjects free–rode the most, and Regular subjects the least. This suggests that subtracting incentives for kindness reduces considerably contribution.

The bottom of Table 4.4 uses the Rank condition by itself and the comparisons between Rank and RegRank and between Regular and RegRank in order to separate cooperation into each of the possible motives for every round. The measures of kindness and confusion reported in such a table indicate that over rounds 1–6 the total amount of cooperation was stable but the amount of confusion was declining rapidly and that of kindness was increasing; after round 6 confusion was rather stable, but kindness fell. Such a pattern points to a possible explanation for the decay phenomenon observed in public goods ex-

periments: "when individuals who start off confused finally learn the dominant strategy, it appears that they may first try to cooperate but then eventually turn to free–riding. This could suggest that, for some subjects, kindness may depend on reciprocity" (Andreoni 1995a, pp. 897–898).

Andreoni's conclusions from this study can be summarised as follows:
– both confusion and kindness are significantly present and merit greater consideration than what received;
– confusion is especially apparent in the experiment because errors can only be in one direction (and so will not be averaged out of the aggregate data);
– the focus on learning in experimental research should shift to include studies of preference for cooperation.

Palfrey and Prisbey (1996) maintained that Andreoni's paper and conclusions were only a partial recognition of the relevance of noise in explaining cooperative behaviour. According to them, "a careful study...designed to precisely measure the relative contribution of each of the various proposed explanations has not yet been carried out... The typical experimental designs do not permit precise measurement of the separate contribution of these diverse effects: altruism, reputation–building, and noise" (Palfrey and Prisbey 1996, p. 410). As a consequence, they proposed an experiment specifically designed to sort out these effects and directly measure the separate contribution of each.

Their study's basic premise was the statistical decomposition of individual behaviour into a systematic component (the *decision rule*) and a residual component (the *noise*, or *error*). In their opinion, in the context of linear voluntary contribution games, attention can be limited to very simple decision rules, called *cutoff decision rules*, in which individuals contribute if and only if their *mrs* between the private good and the public good is less than or equal to some critical value.[28] As for the noise component, Palfrey and Prisbey interpreted it as a random variation (over time) in subjects' *observed* decision rules due to extraneous factors which are impossible to measure.[29] Given this interpretation of the noise, they expected experience to lead to a decrease in noise and considered such a decrease as evidence of learning.

In order to estimate the distributions of decision rules and error rates, Palfrey and Prisbey systematically varied each subject's *mrs*. In every period, the *mrs* were independently drawn, for each individual, from a range of different

[28]Perfectly self–interest behaviour is a special case where the critical value is 1; whilst, altruistic or reputation–building behaviours would be consistent with decision rules where the critical value is set higher than 1.

[29]These factors would include, for instance, computational errors or errors associated with learning by doing.

values, such that for some values contributing nothing to the public good was the dominant strategy, while for other values the dominant strategy was to contribute everything to the public good. The variation of the mrs across periods generates data about decisions in different situations.[30]

Then, to break down the systematic component of decision rules and identify the relative importance of altruism and strategic (reputation–building) behaviour, Palfrey and Prisbey followed Andreoni's (1988) approach and used the Partners–Strangers design. If reputation–building is relevant, then Partners should exhibit decision rules with higher critical points than Strangers. In addition, Partners should show significantly more decay.

Let us now describe the details of Palfrey and Prisbey's experiment and the findings they obtained. They considered group of N subjects, each endowed with a divisible amount X_i of a private good which could be either kept or contributed to a public good. The marginal rate of transformation between the public and the private goods was one–for–one, and individual i's payoffs were generated by the following function:

$$U_i(x_i, x_{-i}) = Vx_i + V\sum_{j\neq i} x_j + r_i(X_i - x_i), \tag{4.5}$$

where x_i is individual i's contribution, V is the marginal value of the public good and it is the same for all individuals, r_i is the marginal value of the private good and it is private information.[31]

By varying r_i $(\forall i \in N)$ over a sequence of 10 decision–periods, Palfrey and Prisbey were able to estimate i's decision rule $D_i(r_i/V)$, where r_i/V is i's mrs. Theoretically, such a decision rule should take the following form:

$$D_i(r_i/V) = \begin{cases} 0, & \text{if } r_i/V < 1 + a_i + \epsilon_i, \\ X_i, & \text{otherwise,} \end{cases} \tag{4.6}$$

where a_i is i's level of altruism and ϵ_i is a random error term. Such a rule represented the *cut–point rule* and the value $c_i = 1 + a_i$ was the *cut–point*. When $\epsilon_i = 0$ and the game has a finite number of periods, rule (4.6) coincides with the dominant–strategy rule. The inclusion of ϵ_i accounts for the possibility of random errors or subjects' unpredictable behaviour.

---

[30] None of the previous experiments varied individual incentives across decisions, nor did they provide explicit information about the distribution of incentives in the population. Palfrey and Rosenthal (1991) used an environment similar to the one explored by Palfrey and Prisbey, but the public good technology was step–level, not linear.

[31] In the experiment, the following parameters' values were chosen: $N = 4$, $X_i = 9$, and $V = 6$ or 10 points according to treatment.

Palfrey and Prisbey estimated subjects' decision rules by using the separate techniques of ordered probit and classification errors analyses.

In the classification analysis, the authors proceeded by determining the rate of classification errors for each possible cut–point. For each token and for each subject i, i's decision was classified as an *error* if, under the hypothetical cut–point rule (4.6), i should have contributed the token but he did not, or i should have not contributed but he did. At the aggregate level, Palfrey and Prisbey estimated the common cut–point c^* as the cut–point with the fewest classification errors. The common error rate ϵ^* was then set equal to the rate of classification errors if c^* is the cut–point.

Data showed that, for both the Strangers and the Partners treatments, $c^* = 1$. This, according to Palfrey and Prisbey, suggests homogeneity of subjects' decision rules and no evidence either for altruism or for reputation–building. However, since in their experiment Strangers exhibited an higher error rate than Partners, the authors maintained that their data support the 'noise' hypothesis as explanation of the differences between the two treatments.

An alternative approach used by Palfrey and Prisbey to measure an 'average decision rule' among subjects was the ordered probit analysis, in which they estimated the probability of any number of tokens contributed as a function of the mrs.[32] The results obtained with such an analysis confirmed those just described. Specifically:

1) more noise in the Strangers treatments than in the Partners ones: strangers had flatter expected contribution curves and, therefore, a higher error rate;[33]

2) reduction of noise with experience: inexperienced subjects exhibited flatter response curves than experienced subjects;

3) effect of 10–period repetition similar to experience effect: the response curves were steeper in the last half of a 10–period session than in the first half, even if the average contribution rate was unchanged. Such a result contradicts past findings of significant decay in contribution rates. But, according to Palfrey and Prisbey, there is no contradiction at all. "It simply means that the observed decay in past experiments was due to learning, not reputation" (Palfrey and Prisbey, 1996, p. 422). This lack of reputation effects was also documented in the less decay found in the Partners treatment than in the Strangers one (although the difference was not statistically significant).

From all these results, Palfrey and Prisbey's concluded that the observed

[32] This analysis implicitly assumes that all subjects use the same decision rule and that the random error terms ϵ_i have a normal distribution with mean zero.

[33] Analysis at individual level then demonstrated that this might be due to more individual errors, more variance across subjects, or a combination of both.

violations of the dominant–strategy prediction in voluntary contribution experiments are due neither to altruism nor to strategic reputation–building behaviour, but to statistical fluctuations in subjects' decisions, manifested as random noise in the data.

Although the amount of noise in the data may explain the observed cooperation, Palfrey and Prisbey's findings appear doubtful for two reasons. First, their choice of the cutoff decision rule may significantly influence the results. Second, the great relevance attributed in the design of the decision problems to the *mrs* may influence players' behaviour, inducing them to concentrate on that aspect.

In a more recent paper, Palfrey and Prisbey (1997) used the same design as in their 1996 paper to separate errors, linear altruism and 'warm–glow'.[34] Also in this study, they focused on two aspects of the data: identifying errors or background noise (namely, observed behaviour which is inconsistent with standard theory) and measuring response functions. They estimated such functions at both the aggregate and the individual levels by using probit models and dichotomising the contribution decision (so that $\forall i \in N$, $x_i \in \{0, 1\}$).

Allowing for warm–glow and linear altruism, the representative individual i's utility derived from the two options contr(ibuting) and not–contr(ibuting) was:

$$\begin{aligned} U_i(\textbf{contr}) = V \sum_{j \neq i} x_j + V + \gamma_i + a_i \sum_{j \neq i} pr_j + \\ + a_i(N-1)V \sum_{j \neq i} x_j + a_i(N-1)V + \epsilon_i \end{aligned} \tag{4.7}$$

$$U_i(\textbf{not–contr}) = V \sum_{j \neq i} x_j + r_i + a_i \sum_{j \neq i} pr_j + a_i(N-1)V \sum_{j \neq i} x_j, \tag{4.8}$$

where V, r_i, and x_i are as in Eq. (4.5), N is the number of players in i's group, γ_i is i's warm–glow term, and pr_j denotes the benefits that subject $j \neq i$ derives from his private account. a_i and ϵ_i are (as in (4.6)) player i's linear altruism and error terms, respectively. Palfrey and Prisbey assumed that the error terms were independent, identical, normally distributed random variables with mean 0 and standard deviation σ. Note that, in the linear altruism model, i's utility

[34] As pointed out on p. 51, the last term—introduced by Andreoni (1989; 1990)—is meant to capture the utility which individuals derive from the act of contributing *per se*.

is a linear combination of his own earnings and the earnings of the other $(N-1)$ members of his group.

Given Eqs. (4.7) and (4.8), the *net utility from contributing* is:

$$U_i = U_i(\textbf{contr}) - U_i(\textbf{not–contr}) = \gamma_i + (V - r_i) + a_i(N-1)V + \epsilon_i. \quad (4.9)$$

Assuming that subject i would contribute if and only if $U_i \geq 0$, this gives the following dichotomous probit model:

$$\epsilon_i \geq (r_i - V) - \gamma_i - a_i(N-1)V, \quad (4.10)$$

where the right–hand side contains all the elements of the subject's utility function which determine his choice x_i. Inequality (4.10) was therefore the decision rule of this model.

First, Palfrey and Prisbey carried out an aggregate analysis. Accordingly, the warm–glow and altruism effects were assumed to be the same across individuals; i.e., $\forall i \in N$, $\gamma_i = \gamma$ and $a_i = a$. Given the specification of the subject's decision rule and utility functions (4.9), by estimating a constant term and the coefficients of the variables $(V - r_i)$ and V, Palfrey and Prisbey obtained estimates of γ/σ, $1/\sigma$ and $a(N-1)/\sigma$, respectively. Thus, through algebric manipulation, they could directly estimate the occurrence of error (σ), warm glow (γ) and linear altruism (a).

Palfrey and Prisbey controlled also the effects of three more variables: *experience*, which took value of 0 for decisions in the first ten–period sequence and 1 for decisions in the second ten–period sequence; *period*, which varied from 1 to 10; and *endow*, which was 0 if the endowment was indivisible and 1 if it was divisible.

From such aggregate probit analysis, Palfrey and Prisbey found strong evidence for warm glow and errors, but not for linear altruism. They also found that experience and repetition were significant explanatory variables,[35] both leading to a decline in contribution rates. But much of this decline was due to a reduction in error rates rather than to a change in the underlying decision rule.

Palfrey and Prisbey further broke down the aggregate analysis at the individual level by including a dummy variable for each individual, from which they estimated the actual distribution of individual warm–glow effects.[36] They found that considerably less than half the subjects had a warm–glow term signif-

[35] The coefficient on the *endow* treatment variable was, instead, insignificant.

[36] Palfrey and Prisbey felt confident that this specification captured the key component of subject heterogeneity in their experiment.

icantly greater than zero, and that no subject exhibited a significantly negative warm–glow term.

Thus, this study's findings support those in Palfrey and Prisbey (1996) by suggesting that the decay phenomenon observed in public goods experiments is the result mainly of a reduction in the amount of subjects' decision errors *combined with* a lower variance in the distribution of individual warm–glow effects. It is not due to an overall decline in warm glow. This implies that "players do not become significantly more selfish with experience";[37] rather, their preferences were shown to be relatively stable with respect to experience.

A further answer to the relative importance of cooperative or altruistic motives and errors in explaining the established contribution result comes from Brandts and Schram (1997). In their attempts "to take the study of voluntary contributions to public good one step further",[38] they proposed a new design for voluntary contribution mechanism experiments, in which subjects were requested to supply a complete 'contribution function' in every period, deciding *a priori* on their contribution level for different possible mrs. The main advantage of this new design is that it yields very rich information about individual behaviour, since in every period subjects made a number of decisions equal to the number of different mrs. Moreover, the values of the mrs were chosen such that for some of them it was a dominant and efficient strategy to contribute everything to the public good, for others it was dominant but not efficient to contribute nothing (the normal dilemma situation), and for others it was both dominant and efficient to contribute nothing.

This design avoids a possible problem of Palfrey and Prisbey's design, namely a kind of instability into the environment induced by their having an independent random drawing for each individual in each period. Nevertheless, it preserves the important advantage of not concentrating exclusively on corner solution situations.

Brandts and Schram's experiment consisted of various sessions, each of which was run twice: once in Amsterdam and once in Barcelona. Twelve subjects (divided randomly in 3 groups of 4) took part in every session that was implemented in a sequence of 10 decision–rounds. The task of the subjects was to invest 9 tokens in any of two accounts: A, the group account (yielding an identical amount of money to every member of the group), or B, the private account (yielding money to the subject alone). In each round, the division had to be made for 10 different situations characterised by different mrs, obtained

[37] Cf., Palfrey and Prisbey (1997, p. 842).

[38] Cf., Brandts and Schram (1997, p. 3).

keeping the payoff to the public account constant but varying that to the private account. After each subject had finalised the division of tokens for all situations (i.e., *mrs*), one of these was randomly selected to be played and the experiment proceeded to the next period.

Three treatments were studied: 1) homogeneous versus heterogeneous situations, 2) partial versus full information and 3) partners versus strangers

The first treatment varied the way in which the *mrs* were selected. In the heterogeneous situation, a *mrs* was chosen separately for each individual (like in Palfrey and Prisbey's experiment) whilst in the homogeneous situation, the monitor selected a *mrs* per group.

The second treatment (partial versus full information) controlled for the amount of information received by the subjects. In the partial information case, only the actual investment in the group account was given;[39] in the full information case, each subject was told the sum of investments in the public account for each of the *mrs* separately.

The last treatment distinguished a Partners case and a Strangers case like in Andreoni 1988.

Brandts and Schram started the presentation of their results with an aggregate data analysis. They estimated two types of contribution functions, both of which were step functions. In particular, they considered a *1–step contribution function* (which starts at any level x of tokens and has at the most a downward step towards a lower level y of tokens)[40] and a *m–step contribution function* (which allows for multiple steps, at any location (i.e., *mrs*) and for any number of tokens between 0 and 9 to be allocated at any *mrs*). For each of the two step functions, that minimising the sum of squared errors was chosen as the one best characterising subjects' aggregate behaviour.

For all treatments together, Brandts and Schram calculated the best step functions over all periods and for period 10 only.

The estimated m–step contribution function for all periods presented four steps; and the one for period 10 exhibited three steps. For the first two *mrs* (when it is a dominant strategy to contribute everything) the value of both these step functions was 9, and for the last two *mrs* (when it is both dominant and efficient to contribute nothing) their value was zero. Therefore, subjects in Brandts and Schram's experiment appeared to recognise the different situations:

[39] Notice that this is not very informative for the heterogeneous situation, since a player did not know on which *mrs* the others' decisions were based.

[40] The 1–step function with $x = 9$ and $y = 0$ is that investigated by Palfrey and Prisbey (1996). They minimised absolute errors, but, in their case, this is equivalent to minimise the sum of squared errors.

(in the aggregate) they did invest all their tokens in the public account when it was a dominant strategy to do so, while they did not put anything in the public account when it was inefficient to do so. In addition, the estimated m–step function overall periods predicted contributions of 12 out of 54 tokens in situations 3–8, where standard theory predicts zero.

As for the class of 1–step functions, within it, the sum of squared errors was minimised when $x = 9$ and $y = 1$, with the downward step at mrs=1. Standard theory would predict a 1–step function with the step from 9 to 0 tokens at mrs=1. This is the optimal 1–step function in Brandts and Schram's data if they set the restriction $y = 0$. So only imposing this restriction, it is possible to argue that subjects behaved according to the theory.

The authors, then, claimed that the m–step function gives a better estimate of their data than the restricted 1–step function. However, from an econometric point of view, this is a difficult claim to prove since Brandts and Schram did not correct for the number of degrees of freedom used.[41]

When all possible combinations of the three treatments were tested, only the main effect of full versus partial information was found to be statistically significant. Apparently, at the aggregate level, giving better information about the (sum of) the others' contributions yielded higher contributions.

By turning the attention to Brandts and Schram's analysis of behaviour across individuals, their approach was to derive a representative step function for each subject and to classify individuals according to the characteristics of their step function.

Specifically, the subjects whose step functions had only one step from 9 to 0 between situations 2 and 3 were called *individualists* since they behave strictly according to the game–theoretic prediction. The subjects who exhibited step functions with more than one step between situations 3 and 8 (implying positive investments in the public account when the dominant strategy would be to invest zero) were classified as *semi–individualists* or *cooperators* depending on the number of tokens contributed. Subjects who deviated from the dominant strategy in situations 9 and/or 10 (implying positive contributions when zero contributions is both dominant and efficient) were defined *altruists*.

Their data showed that the share of individualists (and semi–individualists) was 43.3%; that of cooperators was 31.8%, which became 42.7% by including altruists; that of unclassified was 14.1%.

As for the effects of the three treatments at the individual level, Brandts

[41] According to the authors, the integer values of the steps in the functions made it difficult to come up with a statistical test which would allow them to correct for the number of degrees of freedom.

and Schram expected that they should affect the behaviour of cooperators but not that of individualists (who are supposed to maximise private earnings irrespective of the treatment they are in). Results confirmed these expectations. For the groups of individualists and unclassified, no significant treatment effects were found. For the cooperators, however, three significant effects were detected which can be summarised as follows: in Partners, full information yielded more cooperation than partial information; under full information, Partners cooperated more than Strangers; and in the heterogeneous case, Partners cooperated more than Strangers. These effects seem to imply that cooperators contributed more when it was easier to find out if other cooperators were present in the group.

A further question addressed by Brandts and Schram was whether the presence of multiple steps in the function characterising aggregate behaviour could be interpreted as (tacit) cooperation or as errors. The authors indicated two main reasons why the noise hypothesis could not be a full explanation of their data. First, the errors were not symmetric around mrs=1. Second, a large fraction of individuals consistently (over 10 periods) contributed substantial amount to the public account, while, in an errors interpretation, one would expect these individuals sometimes to contribute much and sometimes to contribute little.

Finally, Brandts and Schram investigated the presence of 'other–regarding' motivations in their data. First, they analysed warm–glow and linear altruism by estimating a model based on that introduced by Palfrey and Prisbey (1997). Then they added *reciprocal altruism* by following the formulation proposed by Levine (1998).[42]

The evidence from the first analysis provided support for both linear altruism and warm–glow. The presence of a positive and significant linear altruism term contrasts with Palfrey and Prisbey's findings. In addition, when Brandts and Schram estimated the warm–glow terms in model (4.10) at the individual level and considered the concentration of warm–glow in groups for periods 1 and 10, they found that such a concentration in specific groups was much larger in period 10 than in period 1. In the standard interpretation, warm–glow is a trait which should not be affected by interaction with others. This, according to Brandts and Schram, suggests that model (4.10) may be misspecified.

Levine's (1998) model allowed the authors to incorporate reciprocal altruism in (4.10). Results from this extended model showed a slight increase in the warm–glow term, a virtually unaltered value for the error σ, and a positive and

[42] In Levine's (1998) model, the weight attributed to the others' income depends on what the others' level of altruism is believed to be. Levine found support for his model in data from various kinds of experiments, including voluntary contribution mechanisms.

significant value of the parameter that brought reciprocity in the model.

Brandts and Schram's experiment makes it evident that the important issue is no longer whether cooperation takes place, but what kind of cooperative model best explains the data. While Palfrey and Prisbey (1997) attributed observed contributions to a combination of warm–glow and error, Brandts and Schram cast some doubt on the warm–glow being the only kind of other–regarding behaviour. Indeed, their experimental results indicated that some kind of reciprocal motivation is necessary to explain the data. This also means that the linear altruism model is not sufficient as a cooperative model. Other aspects of public goods dilemma need, therefore, to be considered among which the role of reciprocal motivations and their weight in comparison with alternative theories of cooperation.

4.3.3 Framing, altruism, reciprocation and distributive concerns

Andreoni (1995b) tried to address the puzzling question of the diversity of outcomes generated by public goods experiments with respect to an important collection of other experiments with externalities, such as common–pool resource and oligopoly experiments, which (unlike the former) tend to confirm the Nash equilibrium prediction. He noticed that one main difference between these two categories of experiments is that in public goods experiments subjects are asked to generate negative externalities while, in all the others, subjects generate positive externalities. Since, in Andreoni's opinion, such a difference alone might explain the gap between the results, he examined the effects of positive and negative framing on cooperation.

Thus, he compared a standard public good game (that he called *positive–frame* condition since, in it, the subject's choice is modelled as contributing to a public good which has a positive benefit to other people) with one obtained by simply framing the subject's choice as purchasing a private good that, since the opportunity cost is investing in the public good, makes the others worse off. This was the *negative–frame* condition which, aside from the different decision framing, preserved the subjects' incentives: in it, a self–interested player still had a dominant strategy to free–ride.

In his experiment, Andreoni used 80 subjects (40 for each condition). Participants played the game for 10 rounds and, in order to avoid reputation building, they were randomly assigned to new groups each round.

Andreoni found that the difference between the two conditions was quite striking and statistically significant, with subjects being much more coopera-

tive in the positive–frame condition than in the other. In particular, average contributions were 33.6% in the positive frame and only 16.2% in the negative frame. The difference, round by round, between the two conditions went from a minimum of 4.9% to a maximum of 26.1%, with an average of 17.4%. A Mann–Whitney rank–sum test rejected the null hypothesis of no difference between conditions at a level of significance beyond the 0.001 percent.[43]

Similar results were obtained with respect to the percentage of subjects who chose the dominant strategy of free–riding during each iteration of the game. On average, 63.5% of negative–frame subjects free–rode in any round, which was nearly twice the rate of positive–frame subjects (34.5%). In addition, the difference between frames increased over the course of the experiment.

The hypothesis that the negative frame made the incentives clearer to the subjects was rejected on the basis of a post–experimental questionnaire. Only one subject in the positive–frame session failed to recognise what choice on his part would maximise his payoffs.

Hence, according to Andreoni's data, framing the choice as a positive externality substantially increases cooperation over framing the decision as a negative externality.[44]

From such a result Andreoni deduced that the positive frame of the game causes certain behaviours that are not activated at the same degree by framing the decision as a negative externality, and that, in order to explain this, it is necessary to go beyond an assumption of pure altruism.[45] Thus, Andreoni's conclusions were that "there must be some asymmetry in the way people feel personally about doing good for others versus not doing bad: the warm–glow must be stronger than the cold–prickle" (Andreoni 1995b, p. 13).

As emphasised earlier, another line of experimental inquiry into subjects' contributing behaviour focused on the role played by reciprocity theories and on their relevance with respect to alternative models of cooperative behaviour.

Bolton, Brandts and Katok (1997) (henceforth, BBK) considered six dif-

[43]This test organises the data by subjects and is normally distributed. It was conducted by first calculating the mean contribution level for each subject and then ranking subjects by these means in the joint sample. Under the null hypothesis, the sum of the ranks should be equal across conditions.

[44]This result was somehow questioned by the evidence reported in Burlando and Hey (1997). As pointed out earlier, in this paper, the authors were not interested in a comparison of framing, but they had a sort of negative frame (different from that of Andreoni) since they rephrased the problem in terms of public bad instead of public good. Their results were in between those registered by Andreoni (1988) and those registered by Weimann (1994), both of whom used a positive frame. Thus, Burlando and Hey's conclusions were that other differences (like nationality) are more relevant than the kind of framing explored by Andreoni.

[45]Since the payoff space was identical in both frames, caring only about the others' payoffs (as pure altruism models state) is not sufficient to generate the differences observed.

ferent hypotheses for contributions in dilemma games. Each hypothesis was classified in two ways: 1) by whether the key concept was altruism, mistakes or norms of cooperation;[46] and 2) by whether the hypothesis was strategy sensitive (in the sense that its predictions about contributions depend on information about others' behaviour).

Table 4.5 summarises the six hypotheses 'constructed' by BBK.

Table 4.5: Classification of BBK's hypotheses by key concept and strategy sensitivity.

	Strategy Sensitive	*Strategy Insensitive*
Altruism	Distributive altruism	Joint payoff maximisation
Norms of Cooperation	Direct reciprocity	Unconditional cooperation
Mistakes	Misidentification	Random errors

Since these hypotheses were tested in a simple two–person dilemma game, I shall use this type of game in order to describe them. For this purpose, let us indicate the two players with A and B, and let x_A and x_B denote their respective monetary payoffs.

Distributive altruism asserts that distributional concerns are central, and it relates to the equity (or inequality aversion) models discussed in Subsection (4.2.2).[47] According to this hypothesis, contributing player A has preferences over the 'commodity bundle' (x_A, x_B), which (by following Becker's (1974) public good model) was interpreted as 'social income'. In order to obtain sharp predictions, it was assumed that both arguments in the utility function were normal goods.[48] Distributive altruism is strategy sensitive because A's contribution depends critically on B's contribution which affects the available social income.

According to the alternative concept of altruism, *joint payoff maximisation*, contributing player A prefers a higher value of $x_A + x_B$ to a lower value, so that only efficiency matters, distribution does not. Such hypothesis is strategy insensitive since contributions are independent on A's expectations of what B will do.

As for the two norms of cooperation, *direct reciprocity* says that player A

[46] As will be clarified later, by norms of cooperation BBK mean those theories which reformulate the choice rule of orthodox game theory and which have been discussed in Subsection (4.2.3).

[47] In such models, indeed, players are assumed to care about the distribution of payoffs between self and the rest of the group.

[48] The latter assumption implies that as the available social income rises, A's choice of both x_A and x_B rises.

contributes more than the minimum only if he expects B to do likewise (so it is strategy sensitive), whilst *unconditional cooperation* states that player A contributes a positive amount independent of what he expects B to do (so it is strategy insensitive).[49]

Finally, the two mistakes hypotheses are the ones identified on p. 58 exactly as *random errors* and *misidentification*. The former asserts that contributing player A chooses his strategy randomly and it is, therefore, strategy insensitive. The latter claims that player A, failing to recognise the dominant strategy, contributes a positive amount because this is the social norm he associates with the everyday game. Misidentification may be strategy sensitive because, by observing B's behaviour, A could learn the equilibrium of the game.

In order to test these hypotheses, BBK exploited their being strategy sensitive or not. Accordingly, they run three different treatments which varied what A knew about B's choice. The *uninformed treatment* was a standard simultaneous–move dilemma game: player A chose a level of contributions between 1 and 6 without knowing B's choice, and player B chose one of two levels of contributions (either 1 or 2) without knowing A's choice. The *informed treatment* differed from the previous one in only one way: player A chose contingent on B's choice; that is, A made two choices, one for each of B's potential contributions.[50] Finally, the *dictator treatment* was as the uninformed treatment, except that, in it, B had only one choice (i.e., contributing 2); this made A a 'dictator'. The dictator game shares two important features with dilemma games: A's dominant strategy is to contribute the minimum possible, and deviations from such a dominant strategy increase joint payoffs.

Let $\hat{g}_A$ be player A's mean contributions. By thinking of the informed treatment as composed of two halves: one for A's choice when B chooses 1 and one for A's choice when B chooses 2, the prediction of each hypothesis about $\hat{g}_A$'s pattern across treatments (i.e., about how $\hat{g}_A$ will (or will not) shift) can be summarised as depicted in Table 4.6.[51]

In the experiment, each subject played in two treatments whose corresponding games were one–shot (i.e., were played just once).[52] A and B roles were

[49] Notice that unconditional cooperation is what I referred to as Kantian behaviour, while direct reciprocity is a simplified version of Sugden's principle of reciprocity.

[50] This game is strategically equivalent to having B choosing first and A choosing after being informed of B's choice. Such a method provides a more complete portrait of A's behaviour. Although, according to standard theory, asking a subject for a conditional choice is equivalent to asking for a choice after he sees his partner's move, these two methods can produce modest but significant differences in results (cf., Camerer et al. (1996)).

[51] See BBK (1997, pp. 12–13) for an explanation of this table's entries.

[52] According to BBK, the simplicity of their game made repetition for the sake of learning about the environment unnecessary.

Table 4.6: BBK's (1997) experiment: Predictions about player A's mean contributions.

Hypotheses	**Predictions**
Joint Payoff Maximisation	$\hat{g}_A^U = \hat{g}_A^{I1} = \hat{g}_A^{I2} = \hat{g}_A^D = 6$
Distributive Altruism	$\hat{g}_A^{I2} = \hat{g}_A^D > \hat{g}_A^U \geq \hat{g}_A^{I1}$
Unconditional Cooperation	$\hat{g}_A^U = \hat{g}_A^{I1} = \hat{g}_A^{I2} = \hat{g}_A^D$
Direct Reciprocity	$\hat{g}_A^{I2} > \hat{g}_A^{I1} = \hat{g}_A^D; \quad \hat{g}_A^{I2} \geq \hat{g}_A^U \geq \hat{g}_A^{I1}$
Random Errors	$\hat{g}_A^U = \hat{g}_A^{I1} = \hat{g}_A^{I2} = \hat{g}_A^D$
Misidentification	$\hat{g}_A^{I1} = \hat{g}_A^{I2}; \quad \hat{g}_A^U \geq \hat{g}_A^D$

U = Uninformed treatment; $I1$ = Informed treatment when B chooses 1; $I2$ = Informed treatment when B chooses 2; D = Dictator

alternated between treatments and subjects were *not* told the outcome of the first game prior to playing the second (thus, the experiment had a 'no feedback' design).

Data analysis showed that contributions were higher in the dictator treatment and in the informed treatment when B chooses 2 than in the other two treatments. Specifically, in terms of Table 4.6 notation, the experimental results can be stated as follows:

$$\hat{g}_A^{I2} = \hat{g}_A^D > \hat{g}_A^U = \hat{g}_A^{I1}. \tag{4.11}$$

BBK, then, considered which of the six hypotheses had the reasonably accurate predictive power so as to explain the results summarised in (4.11). They found that all strategy insensitive hypotheses as well as misidentification failed to capture one (or more) aspect(s) of the observed pattern of comparative statics, and that direct reciprocity failed not only in capturing observed features but also because some of its predictions were not detected. Only distributive altruism appeared to be both sufficient and necessary to explain the data. Sufficient because by itself it was consistent with (4.11); i.e., it matched the observed pattern of contributions. Necessary in that no combination of the remaining five hypotheses was consistent with (4.11).

Their findings led the authors to two conclusions. First, an explanation of contributing behaviour must be strategy sensitive. Second, out of the strategy sensitive hypotheses, only distributive altruism seems to be an appropriate basis for modelling contributing behaviour. Thus, BBK's study—suggesting a role for distributional considerations—supports equity theories as a plausible expla-

nation for cooperative behaviour; these theories affirm, indeed, that it suffices to consider preferences over payoff allocations for explaining cooperation.

The lack of any evidence for direct reciprocity appears a quite surprising result. The high contributions of player A in the dictator game (where B does nothing) demonstrated that not all contributions can be attributed to rewarding partners for doing the same. On the other hand, the failure to observe higher A's contributions in the informed treatment when B chooses 2 (and so he performs a worthy act) confirmed that the proportion of subjects behaving in accordance with reciprocal motivations was negligible. Nevertheless, there may be some doubts about the power of the statistical tests used by BBK to prove such a result.[53] In addition, the fact that the informed treatment was divided in two parts and the tests were conducted separately for each part might have influenced the results by hiding *some* direct reciprocity.

Bolton, Brandts and Ockenfels (1998) extended the experiment of BBK to a test of two hypotheses, both of which were considered as two characterisations of reciprocity: the *intentional hypothesis* and the *distributional hypothesis.* The former asserts that reciprocity is triggered by well or ill–intended acts so that the reciprocal agent must interpret whether an action was intended to help or hurt him; once determined, he responds with what he considers a fair return. The alternative characterisation of reciprocity is on the line of the equity theories and says that the reciprocator simply acts to implement his exogenous preferences over payoff distributions, independently of any assessment of intentionality. Results of this experiment confirmed those of BBK by providing strong evidence for distributional preferences and, thus, for equity theories.

However neither the ERC model nor the Fehr–Schmidt model can explain results such as Blount (1995) and Charness and Haruvy (1999).

Blount's (1995) experiment looked for evidence of equity in two ways. Second mover rejections in a standard ultimatum game were compared first with rejections in a treatment in which an outside party, receiving no payoff for the game, named the proposal, and second with rejections in a treatment in which the proposal was randomly selected. Applied in the most straightforward manner, equity models predict the same rejection rate in all three treatments. The experiment contradicted this prediction: the 'minimum acceptable offer' averaged 12% of the pie in the random treatment, but 29% in the outside party treatment.

---

[53] In particular, the authors statistically compare contribution distributions in the various treatments by using a t–test of the difference between means (which checks for a location shift) and an empirical χ^2–test whose p–values were averages from five 20.000 trial samplings of the contingency table distribution.

Charness and Haruvy (1999) applied Blount's framework to a gift–exchange game with employers and employees,[54] where the source of the wage differed across treatments and was known to be generated or by a self–interested employer (the standard treatment), or by a draw from a bingo cage (the random treatment), or by the experimenter in advance (the outside party treatment). The authors' aim was to assess the relative success of three approaches to cooperative behaviour (altruism, equity and intentional reciprocity) in explaining their data. They found that equity models outdid pure altruism explanations when employers determined wage, but that the former were inferior to the latter when wages were determined exogenously. Thus, "... it seems that, despite the important insight they provide, neither the simple altruism nor the pure equity–based models can explain behavior in the experiments analyzed in this paper" (Charness and Haruvy 1999, p. 25). On the other hand, also the conventional model of reciprocity (which does not explicitly address distributive concerns) was found unsuccessful, while a formulation of reciprocity which nests altruism as a special case did well. Such results led the authors to conclude that "any successful model must accommodate the concerns of altruism, distribution and reciprocity" (Charness and Haruvy 1999, p. 28).

A further experiment directed to distinguish between alternative behavioural theories was carried out by Croson (1998b), whose results contrasted with those of Bolton, Brandts and Katok (1997), and Bolton, Brandts and Ockenfels (1998) since they provided strong support for reciprocity theories over either theories of altruism and commitment.

She run four separate experiments designed to distinguish among these theories by comparing their statics predictions. In all experiments, she used a typical voluntary contribution mechanism environment as that described in Subsection (2.3.2), where $N = 4$, $m = 25$, and MPCR=0.5. The mechanism was implemented in two sequences of 10 rounds each. Subjects played the first 10–period game, and then were told that there was enough time to play a second identical 10–period game (as in Andreoni (1988)).[55]

In the first of her four experimental studies, Croson tested the comparative

---

[54] A classical experiment involving the gift–exchange game was proposed by Fehr et al. (1993). In this experiment, subjects assigned the role of firms offer a wage to those assigned the role of workers. The worker who accepts the wage then chooses an effort level. The higher the level chosen, the higher the firm's profit and the lower the worker's payoff. The game is essentially a sequential prisoner's dilemma, in which the worker has a dominant strategy to choose the lowest possible effort. The only subgame–perfect wage offer is the reservation wage.

[55] Doubts about the no announcement of the restart from the beginning of the experiment have been expressed on p. 64 of this work, with respect to Andreoni's (1988) study.

static predictions of models of commitment, altruism and reciprocity by investigating the relationship between an individual's contributions and his beliefs about the contributions of the others.

Accordingly, in such a study, people's expectations about others' behaviour were elicited. Specifically, each period, before taking a contribution decision, subjects were asked to estimate the total number of tokens that the other three group members would contribute to the public good in the upcoming decision period. Subjects were compensated for accurate estimates.

Given Croson's interpretation of the three models,[56] in her experiment, models of commitment predict a zero relationship between subjects' contributions and their expectations, models of altruism predict a negative relationship and models of reciprocity a positive relationship.

Random effects regressions estimated either over all 20 periods or by separating the first 10 periods from the second 10 periods of the game yielded identical results. Both reported a significant positive relationship between an individual's expectations and his own contributions. Such a result strongly supported reciprocity theories over the alternative two models.

An analysis carried out at the individual level (by calculating the interesting relationship for every participant) revealed that 22 out of 24 subjects (almost 92%) exhibited a positive relationship, consistent with models of reciprocity; only two subjects presented a negative relationship; and no–one a zero relationship.

Although all these findings appeared encouraging for reciprocity theories, it might be *either* that asking subjects to estimate the others' contributions led them to think reciprocally where they would not otherwise (an 'elicitation hypothesis'), *or* that the repeated–game nature of the experiment caused the positive relationship rather than reciprocity *per se* (a 'reputation hypothesis'). In order to verify these two possibilities, Croson run two more experiments for control. Neither of these further experiments involved the elicitation of expectations about others' action.

Except for the exclusion of the estimation stage, the experiment directed to test the elicitation hypothesis was identical to the first one. Twenty–four subjects, different from the previous subjects but from the same subject pool, participated in this experiment, arranged in 6 groups of 4.

Here, Croson was interested in the relationship between individuals' own contributions and the *actual* total contributions of the other group members. A zero relationship is predicted by commitment theories, a negative relationship

[56] She adopted the models of altruism, reciprocity and commitment presented in Subsections (4.2.2) and (4.2.3).

by altruism theories, and a positive relationship by reciprocity theories.

Random effects regressions estimated for both the previous and this new experiments reported, in accordance with the reciprocity hypothesis, a significant positive relationship. Analysis at the individual level showed that, in the previous experiment, 21 out of 24 subjects exhibited a positive relationship between their own contribution and the actual contributions of the others, only 3 exhibited a negative relationship and no subject exhibited a zero relationship. The pattern in the experiment without the guessing stage was similar: 19 out of 24 subjects exhibited a positive relationship, only 5 a negative relationship and no–one a zero relationship.

Therefore, even if there were some differences in the level of contributions between the two studies, results from this second experiment demonstrated that the comparative statics of reciprocity models remain the most consistent with the data even when expectations are not elicited. Such a result allowed Croson to reject the elicitation hypothesis.

As pointed out earlier, a further concern of the author was that the observed positive relationship arose from some sort of reputation–building rather than from reciprocity. Hence, to directly test this reputation hypothesis, Croson run a third experiment, in which she implemented Strangers sessions: after each period, subjects were randomly re–assigned to new groups of four. If the previously observed positive relationship was due to reputation issues, one should observe a zero relationship in the experiment with Strangers.

At odds with the reputation hypothesis and consistent with reciprocity models, a significant positive relationship between own and others' actual contributions was detected in this experiment as well. In addition, by comparing the interesting relationship at the individual level, Croson found that most of the subjects (almost 71%) exhibited a positive relationship. In contrast to the previous experiments, however, 4 subjects displayed a zero relationship; closer inspection of the data revealed that such relationships were generated by subjects who fully free–rode (i.e., contributed zero) throughout the entire experiment.[57]

Thus, while there were more free–riders and lower contributions in this experiment than in the previous ones, results were still supportive of the comparative statics of reciprocity theories over those of commitment and altruism.

Having demonstrated support for reciprocal concerns in three different settings, Croson turned to a fourth study in which she provided a characterisation

[57]This result of more free–riding in Strangers experiments than among stable groups of Partners appears consistent with Croson's (1996) previous experimental research. On this topic, see also Keser and van Winden (2000).

of the *type* of reciprocity which individuals exhibit.

As stated in Subsection (4.2.3), Sugden's model of reciprocity suggests that agents will match the *minimum* contribution of the others. In contrast, we can imagine different types of reciprocity in which subjects try to match the average contribution of the others, or even the maximum. Croson (1998b) specifically addressed this question by carrying out an experiment where she distinguished among these different specifications of reciprocity.

In this further experiment, Partners treatments were used, no elicitation of beliefs was made, and all parameter values were the same as in the previous experiments. However, in contrast to the previous experiments, after each period, subjects were informed not only of the aggregate contribution of the other three group members, but also of their *individual* contributions. Thus, subjects could attempt to match the maximum, the minimum or the middle contribution of their fellow members.

Data from this study demonstrated that middle reciprocity was a better predictor than either minimum or maximum reciprocity, which means that subjects tried to match the median or average contributions of the others rather than the minimum (as Sugden suggests) or the maximum.

Croson's (1998b) experimental studies, therefore, indicate reciprocal concerns as the main factors which motivate people to make voluntary contributions in social dilemma situations. According to her data, players act as though part of their objective is to match the contributions of the other group members.

4.4 Summary and conclusions

An overall look at the experimental literature taken up in the previous section reveals that in some cases the results are in quite sharp contrast. The experiments dealing with the issues of learning and experience (with which I began the discussion) found evidence for strategic and non–standard behaviour, but not for simple learning.

In an experiment expressly designed to separate the hypothesis that cooperation is due to kindness, altruism or warm–glow from the hypothesis that it is simply the result of errors or confusion, Andreoni (1995a) found out that on average half of all cooperation comes from subjects who understand free-riding but choose to cooperate out of some form of kindness.

By allowing for interior solutions, Palfrey and Prisbey (1996, 1997) attributed the observed contribution and the decay phenomenon to a combination of warm–glow and error (rather than to altruistic or strategic behaviour).

However, in a related model, Brandt and Schram (1997) cast doubt on the warm–glow being the only kind of non–selfish motivation: while confirming that linear altruism is not sufficient as cooperative model, their study indicated that reciprocal concerns matter as well.

On the other hand, in a simple two–person dilemma game, Bolton, Brandts and Katok (1997), and Bolton, Brandts and Ockenfels (1998) did not find evidence for direct reciprocity but for distributive altruism (preferences over payoff distribution consistent with Becker's utility functions for public goods games). Instead, in a typical voluntary contribution mechanism environment, Croson (1998b) provided strong support for reciprocity theories over theories of altruism or of commitment.

These divergences among results and interpretation of them imply that the issue of which approach best explains the data remains open.[58] Each of the models of cooperative behaviour considered can, in fact, explain outcomes in only some subset of the various experimental settings. No model has shown itself to be able to account for all observed cooperation. This suggests that some of the reasons for people voluntarily contributing to public goods remain to be found.

Thus, the challenge is to construct a model of cooperative behaviour (from principles consistent with the capability and the limitations of the human mind) which can explain previously inexplicable cooperation. This will be the purpose of Chapter 5, where I will propose a new explanation of what motivates people under different circumstances including voluntary contributions to public goods.

The basic premise of this study is that I am not looking for one grand theory capable of explaining all significant phenomena in the games we are interested in. Indeed, in my opinion as well as in that of other authors,[59] many factors influence human behaviour and it is (probably) impossible to capture all of them in a single model.

---

[58] We must, however, take into account the different features of the designs as well as the different purposes which the various experiments pursued.

[59] See, for example, Bolton (1998); and Charness and Haruvy (1999).

Chapter 5

The persuasion hypothesis: what it is and how it works

5.1 Introduction

In this chapter, I shall propose a new alternative hypothesis about what motivates people to voluntarily contribute.

The hypothesis is relatively simple. It rests on two principles that seem to have a strong common sense. An agent understands the benefits that *he himself* can obtain if the public good is provided and, consequently, he contributes in order to *persuade* his partners to perform such an action. Indeed, in a repeated game where verbal communication is forbidden, the only way that a player has got to signal information to the others is by his choices of moves as the game proceeds. Thus, if a subject grasps the advantages that he (himself) can have if all group contribute to the public good, he may decide to contribute a constant amount along a sequence of decision periods in order to set an example to his 'selfish' fellow members and push them towards his same choice.

The fundamental assumption upon which I build my behavioural hypothesis is that, at the start of the game, a persuasive player must be *sure* that from a certain period onwards (let us say from period μ onwards) his selfishly–maximising partners will follow his contributing decision. Only if an agent holds this expectation, it makes sense for him to *unconditionally* contribute (i.e., to contribute independently of the others' behaviour) for the first μ periods of the game.

What happens then if, in the μth period, the persuasive agent realises that the other group members are actually not willing to modify their attitude and keep on free–riding? It seems reasonable to suppose that, if his expectations are not fulfilled, he gives up his constant contribution and starts following an

alternative decision rule.

Specifically, I assume that, in all periods after μ, the persuasive player *reciprocates* the others' decisions. Such a reciprocal behaviour is here taken to mean that, from period $\mu + 1$ forward, he perseveres with his constant contribution as long as he observes that everyone else does so; otherwise, he modifies his contribution in the direction of the others' *average* contribution in the previous period. This implies that he increases (decreases) his contribution if it was below (above) the average of the others.[1]

Two key observations from former dilemma games experiments have formed the basis for the development of my ideas.

First, the frequency with which an individual chooses to cooperate after several cooperative outcomes have occurred. Such a frequency implies that the individual does not take advantage of the others' willingness to cooperate by switching to the immediately rewarding (defecting) strategy. This seems to support a theory of reciprocity (according to which an individual's regard for the utility of the others depends on how 'kind' the others are towards him).

But, if reciprocity is the principle underlying behaviour, it is not easy to explain a second key observation; namely, that people tend to repeat cooperative choices after they have just cooperated *without reciprocation*. Experimental evidence shows, indeed, that there exist subjects who cooperate even if their group members do not do the same. Pure altruism might, probably, account for this result. However, in contrast with altruism theories, previous experimental findings report that cooperative subjects change their attitude towards the others if they refuse to modify their selfish behaviour.

An hypothesis as that proposed in this chapter—under which, a subject cooperates unconditionally for μ periods by expecting that, in $t = \mu$, the others will follow his behaviour, and then he reciprocates the others' observed decisions (so that he will stop cooperation if his expectations are not fulfilled)—appears to fit the reported observations quite well.

5.2 The persuasive player: an automaton with a bit of rationality

Let us define the period after which a player expects that all the others will follow his contributing decision as the *reciprocity–test–period* of a persuasive

[1] Keser's (1997) and Croson's (1998b) experimental studies addressed explicitly the question of which type of reciprocity individuals exhibit. Both authors found that behaviour is oriented towards the average (rather than the minimum or the maximum) contribution of the other group members in the previous period.

agent. Previous section's description and introduction of my hypothesis should make it evident that, given this kind of expectation, a persuasive player can be thought of as an *automaton* who, *during the game*, behaves according to a predefined set of instructions. These require him to contribute a constant amount until period μ, and to switch to a reciprocal rule thereafter.

Although he acts on the basis of programmed instructions during the game, the persuasive agent is allowed to have some kind of rationality at the outset of the game. Indeed, in period 1, he is required to solve a maximisation problem: he must compute his optimal level of contribution, $\bar{g}_p$, on the assumption that the other players will each contribute the Nash equilibrium amount until period $(\mu - 1)$, and will thereafter continue to do so unless he plays $\bar{g}_p$ throughout, in which case they will contribute $\bar{g}_p$.

5.2.1 The persuasive player's rational calculus

Let us consider a very general voluntary contribution mechanism decision environment.

Let us suppose that in the model there are N individuals who interact for T periods, and a single public good. In any one period t, each individual i is endowed with income $m_{i,t} = m$ ($\forall i$ and $\forall t$), which can be either privately consumed or invested in the public good. This is produced by the linear technology: $y_t = \sum_{j=i}^{N} g_{j,t}$.

In every period, the utility of each individual i is an increasing function of the quantity of public good provided, y_t, and a decreasing function of his own contributions, $g_{i,t}$:

$$U_{i,t} = U(g_{i,t}, y_t) \quad \forall i \in N \quad \text{and} \quad \forall t \in [1, T]$$

where $\partial U_{i,t}/\partial g_{i,t} < 0$, and $\partial U_{i,t}/\partial y_t > 0$. Let us indicate with g^* the symmetric fully–rational Nash equilibrium of the game.[2]

In such an environment, the maximisation problem faced by persuasive agent i can be modelled as follows:

$$\max_{g_{i,1}} \quad \sum_{t=1}^{T} U(g_{i,t}, y_t)$$

[2]Notice that the equilibrium can be taken to be symmetric since all individuals have identical utility functions and initial endowments.

subject to:

$$g_{i,t} = g_{i,1} \quad \forall t \in [1, T], \tag{5.1}$$

$$\sum_{k \neq i} g_{k,t} = (N-1)g^* \quad \forall t \in [1, \mu - 1], \tag{5.2}$$

and

$$\sum_{k \neq i} g_{k,t} = (N-1)g_{i,1} \quad \forall t \in [\mu, T]. \tag{5.3}$$

Constraint (5.1) reflects the persuasive agent's conviction to contribute always the same amount. Constraints (5.2) and (5.3) specify the initial beliefs that the persuasive agent entertains about the behaviour of each other member k of his group throughout all game.

By substituting the constraints into the objective function, i's decision problem is equivalent to:

$$\begin{aligned} \max_{g_{i,1}} \quad & V(\mu, g_{i,1}) = \\ & = (\mu - 1) \times U(g_{i,1}, g_{i,1} + (N-1)g^*) + \left[T - (\mu - 1)\right] \times U(g_{i,1}, Ng_{i,1}). \end{aligned} \tag{5.4}$$

Differentiating with respect to $g_{i,1}$ and solving yields the amount $\bar{g}_p$ that the persuasive player contributes constantly for μ periods, before starting reciprocating the others' behaviour. That is:

$$\bar{g}_p(\mu) \in \underset{g_{i,1}}{\text{argmax}} \quad V(\mu, g_{i,1}).$$

Given a (known) utility function, problem (5.4) implies the existence of a relationship between $\bar{g}_p$ and μ, which allows us to derive testable restrictions.

Consider, for instance, a 2–person game with:

$$U(g_{i,t}, y_t) = (m_{i,t} - g_{i,t})y_t \quad \forall i = 1, 2 \quad \forall t \in [1, T] \tag{5.5}$$

where $y_t = g_{i,t} + g_{j,t}$.

The reaction function of player i, here, is:

$$g^*_{i,t} = \frac{1}{2} m_{i,t} - \frac{1}{2} g^*_{j,t}.$$

Let us assume symmetry at $m_{i,t} = m_{j,t} = 1$, so that the equilibrium is

$g^* = 1/3$ and the social optimum is $\hat{g} = 1/2$.

Utility function (5.5) implies the following optimisation by the persuasive agent:

$$\max_{g_{i,1}} \quad (\mu - 1) \times (1 - g_{i,1})(g_{i,1} + 1/3) + \left[T - (\mu - 1)\right] \times (1 - g_{i,1})2g_{i,1}.$$

The solution of this problem is:

$$\bar{g}_p = \frac{T - \frac{2}{3}(\mu - 1)}{2T - (\mu - 1)}. \tag{5.6}$$

It can be seen that if $\mu = T + 1$ (i.e., if the player believes that the others will never be persuaded to contribute $\bar{g}_p$), then the optimal choice is $\bar{g}_p = 1/3$, which is the Nash equilibrium. At the other extreme, if $\mu = 1$ (i.e., if the player expects that his partners will be somehow instantly persuaded), then the optimal choice is $\bar{g}_p = 1/2$, which, in this symmetric case, corresponds to the unique social optimum. For values of μ between 1 and $T + 1$, the solution $\bar{g}_p$ lies strictly between 1/3 and 1/2. Therefore, the higher is the persuasive player's reciprocity–test–period, the lower is the amount that he chooses to contribute until then.

Eq. (5.6) represents a testable restriction insofar as we know the player's reciprocity–test–period: given his μ, if the agent is found contributing an amount different from that determined by (5.6), we can infer that he is not solving maximisation problem (5.4) and, therefore, that he is not following a persuasive strategy.

Since it is not easy to obtain information about the value of μ, more useful in this respect is the inverse of Eq. (5.6):

$$\mu = T\frac{(1 - 2\bar{g}_p)}{\frac{2}{3} - \bar{g}_p} + 1, \tag{5.7}$$

which gives an estimation of the persuasive agent's reciprocity–test–period.

Thus, if an agent is observed to contribute some constant amount $\bar{g}_p$ for more periods than those given by (5.7), without being followed by each of his group members, then we can infer that he is not playing according to the persuasion hypothesis.

Consider, as a further example, a symmetric 3–person game with a Stone-Geary utility function of the form:

$$U(g_{i,t}, y_t) = (w - g_{i,t} - a)^{\alpha}(y_t + b)^{1-\alpha} \quad \forall i = 1, 2, 3 \quad \forall t \in [1, T], \tag{5.8}$$

where $y_t = \sum_{j=1}^{3} g_{j,t}$, and w is an underlying level of income greater than $(g_{i,t} + a)$, and therefore greater than i's endowment m.[3]

Eq. (5.8) can be alternatively expressed as:

$$\ln U(g_{i,t}, y) = \alpha \ln(w - g_{i,t} - a) + (1 - \alpha) \ln(y_t + b).$$

In this symmetric case, the (fully–rational) Nash equilibrium is:

$$g^* = \frac{(1-\alpha)(w-a) - \alpha b}{1 + 2\alpha} \tag{5.9}$$

so that for $0 < \alpha < 1$ and for $-\frac{\alpha}{1-\alpha}(3w + b) < a < w - \frac{\alpha}{1-\alpha}b$, one obtains $0 < g^* < w$.[4]

Furthermore, one can identify a (symmetric) social optimum as that $\hat{g}$ which maximizes $\alpha \ln(w - g - a) + (1 - \alpha) \ln(3g + b)$. Thus:

$$\hat{g} = \frac{3(1-\alpha)(w-a) - \alpha b}{3}. \tag{5.10}$$

Notice that for the values of α and a specified before, and (in addition) for $b > 0$, we have $0 < g^* < \hat{g} < w$.[5]

Let us consider now persuasive agent i's maximisation problem. Here, this is:

$$\begin{aligned} \max_{g_{i,1}} \quad & (\mu - 1) \times [(w - g_{i,1} - a)^{\alpha}(g_{i,1} + 2g^* + b)^{1-\alpha})] + \\ & + [T - (\mu - 1)] \times [(w - g_{i,1} - a)^{\alpha}(3g_{i,1} + b)^{1-\alpha})]. \end{aligned}$$

With such a problem, the equivalent of (5.6) cannot be explicitly obtained. Nevertheless, we can derive a testable restriction—equivalent of (5.7)—which

[3]Thanks to this last assumption we can never have: $(w - g_{i,t} - a) < 0$ in Eq. (5.8).

[4]Since $0 < \alpha < 1$, the denominator of (5.9) is positive. Therefore, $g^* > 0$ if $(1 - \alpha)(w - a) - \alpha b > 0$, which implies $a < w - \frac{\alpha}{1-\alpha}b$.

On the other hand, we have $g^* < w$ if $(1 - \alpha)(w - a) - \alpha b < [1 + 2\alpha]w \Longrightarrow (1 - \alpha)a > -3\alpha w - \alpha b$; solving the last inequality yields $a > -\frac{\alpha}{1-\alpha}(3w + b)$.

[5]Specifically, $\hat{g} > g^*$ if $[3(1-\alpha)(w-a) - \alpha b][1+2\alpha] > 3[(1-\alpha)(w-a) - \alpha b]$, which implies $[3(1 - \alpha)(w - a) - \alpha b] > -\frac{2}{3}b$. If $b > 0$, then the right hand side of the latter inequality is negative. Hence, $\hat{g} > g^*$ whenever the left hand side of the inequality is greater or equal to zero; i.e., whenever $3(1 - \alpha)(w - a) \geq \alpha b$ that yields to have:

$$a \leq w - \frac{\alpha}{1-\alpha}\frac{b}{3}. \tag{5.11}$$

If the parameter a is set such that $g^* > 0$, it must be $a < w - \frac{\alpha}{1-\alpha}b$. Provided that $b > 0$, the latter inequality implies inequality (5.11).

takes the following form:

$$\mu = T\frac{B^{\alpha}[\alpha C - 3(1-\alpha)A]}{C^{\alpha}[(1-\alpha)A - \alpha B] + B^{\alpha}[\alpha C - 3(1-\alpha)A]} + 1, \qquad (5.12)$$

where $A \equiv w - \bar{g}_p - a$, $B \equiv \bar{g}_p + 2g^* + b$, and $C \equiv 3\bar{g}_p + b$.

Knowing the parameters α, a and b, Eq. (5.12) allows us to calculate the value of μ corresponding to each level of contribution. Thus, if an agent is observed to contribute some constant amount $\bar{g}_p$ for more periods than stated in (5.12) without being followed by each of his partners, then we can infer that he is not pursuing a persuasive strategy.

It can be shown that, also in this case, the relationship between μ and $\bar{g}_p$ is inverse. To prove this, instead of using the rather complicated expression (5.12), we can proceed along the following reasoning lines.

We know that, unless $V(\mu, \bar{g}_p) \geq V(\mu, g^*)$, the amount $\bar{g}_p > g^*$ can never be optimal. Thus, a necessary condition for $\bar{g}_p$ to be optimal is that:

$$(\mu - 1)A^{\alpha}B^{1-\alpha} + [T - (\mu - 1)]A^{\alpha}C^{1-\alpha} \geq T(w - g^* - a)^{\alpha}(3g^* + b)^{1-\alpha},$$

from which:

$$\mu \leq T\frac{A^{\alpha}C^{1-\alpha} - (w - g^* - a)^{\alpha}(3g^* + b)^{1-\alpha}}{A^{\alpha}C^{1-\alpha} - A^{\alpha}B^{1-\alpha}} + 1. \qquad (5.13)$$

Now, notice that $A^{\alpha}C^{1-\alpha} \equiv (w - \bar{g}_p - a)^{\alpha}(3\bar{g}_p + b)^{1-\alpha}$, and that the latter expression is the persuasive player's single–period utility when he, together with each of his partners, plays $\bar{g}_p$, i.e. $A^{\alpha}C^{1-\alpha} \equiv U(\bar{g}_p, 3\bar{g}_p)$. On the other hand, $A^{\alpha}B^{1-\alpha} \equiv (w - \bar{g}_p - a)^{\alpha}(\bar{g}_p + 2g^* + b)^{1-\alpha}$, which is the persuasive player's single–period utility when he plays $\bar{g}_p$ and both the others play g^*, i.e., $A^{\alpha}B^{1-\alpha} \equiv U(\bar{g}_p, \bar{g}_p + 2g^*)$. By definition, a decrease in $\bar{g}_p$ leads $U(\bar{g}_p, 3\bar{g}_p)$ to decrease, and $U(\bar{g}_p, \bar{g}_p + 2g^*)$ to increase. As a consequence, *ceteris paribus*, the fraction on the right–hand–side of (5.13) will increase.[6] Given T, this induces a rise in μ.

Consider, finally, a symmetric 3–person game with a linear utility function of the form:

$$U(g_{i,t}, y_t) = r(1 - g_{i,t}) + vy_t \quad \forall i = 1, 2, 3 \quad \forall t \in [1, T]. \qquad (5.14)$$

Let us assume that, $\forall i$ and $\forall t$, $g_{i,t} \in \{0, 1\}$, and that the parameters r and v satisfy the constraints: $r > v$ and $3v > r$. Given these inequalities, the unique

[6] The fraction denominator's decrease (caused by the decrease in $A^{\alpha}C^{1-\alpha}$ as well as by the increase in $A^{\alpha}B^{1-\alpha}$) is, indeed, greater than the numerator's decrease (due only to the decrease in $A^{\alpha}C^{1-\alpha}$).

subgame perfect equilibrium is $g^* = 0$ and the unique social optimum is $\hat{g} = 1$.

In such a game, the problem that a persuasive player faces in the first period is to choose either to cooperate or to defect (i.e., to set either $g_{i,1} = 1$ or $g_{i,1} = 0$, respectively) on the assumption that his partners will each defect until $(\mu - 1)$ and they will continue to do so unless he cooperates throughout, in which case they will cooperate from periods μ to T. Thus, it will pay him to cooperate (rather than defect) if and only if:

$$(\mu - 1)v + [T - (\mu - 1)]3v \geq Tr,$$

which yields:

$$\mu \leq T\frac{3v - r}{2v} + 1. \tag{5.15}$$

.

Eq. (5.15), as previous Eqs. (5.7) and (5.12), is a testable restriction: if, in a game with utility function (5.14), an agent is observed to cooperate for more periods than those given by (5.15), without being followed by each of his partners, this would mean that he is not a persuasive type.

5.2.2 Minimum periods of cooperation necessary for identifying a persuasive player

In this subsection, I will show that, in order to identify an individual as a persuasive player, he must cooperate for at least three rounds. Distinguishing between a persuader and other players' types would be, otherwise, hard.

Let us suppose, indeed, that, in one of the games analysed in the previous subsection, player i contributes a constant amount $\bar{g}_i$ greater than the equilibrium, g^*, only for the first two periods, and that in the meantime (i.e., in $t = 1, 2$) $g_{k,t} = g^*$, $\forall k \neq i$. If player i decreases his contribution in the third period so that $g_{i,3} = g^*$, then i may be a strategic player (as defined in Chapter 4, Subsection (4.2.3)), or he may be following a tit–for–two–tat strategy (according to which a subject starts playing by cooperating twice, and then either he perseveres with cooperation or he shifts to defection depending on the others' decisions). If, instead, *ceteris paribus*, $g_{i,3} = \bar{g}_i$, then the tit–for–two–tat strategy cannot any longer provide reasons for i's behaviour, neither can the reputation building/strategies hypothesis. On the basis of the assumption made in Chapter 4, a strategic player stops being uncertain about the rationality of his partners (and stops, therefore, his pretension to be a cooperative type) whenever he observes the others to play the dominant strategy for two

consecutive periods. This implies that the strategies hypothesis cannot explain the behaviour of a player who does not give up his constant contributions in $t = 3$ even if he has observed the others to free–ride throughout. Thus, only some other explanation can give reason of the fact that a player keeps on his contributing behaviour in $t = 3$ when his partners are free–riding. Such an alternative explanation, I suggest, can be found in a player's willingness to persuade the others to follow him.

5.2.3 The relationship between a persuader's contribution and the others' behaviour

Let $g_{p,t}$ be the contribution of a persuasive player in period t. Then, on the basis of my previous analysis, $g_{p,t}$ must be:

1) independent of the others' behaviour until period μ (where $\mu \geq 3$); in fact, $\forall t \in [1, \mu]$: $g_{p,t} = \bar{g}_p(\mu) \in \underset{g_{i,1}}{\text{argmax}} \quad V(\mu, g_{i,1})$;

2) positively related to the others' behaviour thereafter; in fact, $\forall t \in [\mu + 1, T]$, the persuader must reciprocate his partners' contribution. According to the definition of reciprocity given in the Introduction of this Chapter, this implies *either* that $g_{p,t}$ remains constant (if, $\forall k \neq p$, $g_{k,t-1} = \bar{g}_p(\mu)$); *or* that $g_{p,t}$ changes in the direction of the other group members' average contribution in the previous period (which means that $g_{p,t}$ decreases or increases depending on whether it was, respectively, above or below the average of the others).

In this case, therefore, the relationship between a persuader's contribution and the behaviour of his partners can be described by the following two derivatives:

$$\forall t \in [1, \mu] \quad (\mu \geq 3): \qquad \frac{\partial g_{p,t}}{\partial \sum_{k \neq p} g_{k,t-1}} = 0; \tag{5.16}$$

$$\forall t \in [\mu + 1, T]: \qquad \frac{\partial g_{p,t}}{\partial \sum_{k \neq p} g_{k,t-1}} > 0. \tag{5.17}$$

Relationships (5.16) and (5.17) can be seen as additional testable restrictions: by regressing the contributions of a subject on the total contributions of his partners in the previous period, if the slope of the regression line does not change (from zero to positive) in period μ, then we can infer that the subject is not following a persuasive behaviour.

Let us consider now a 3–person game where each individual is characterised by the linear utility function (5.14), and where, $\forall i$ and $\forall t$, $g_{i,t} = \{0, 1\}$. Given

the binary dimension of the strategy set, a definition of reciprocity with respect to the average of the others' contributions is misleading: if player i cooperates in $t = \mu$ and only one of his fellow members reciprocates him, then, if i is a persuasive player, (according to the above definition of reciprocity) $g_{i,\mu+1}$ should be equal to 0.5, which is not a feasible strategy.

On the other hand, a definition of reciprocal behaviour that requires an individual to keep on cooperation if only one of the other two players cooperates might sound overly restrictive. It appears more plausible to allow a reciprocator to switch to defection as soon as he realises that one of his partners is defecting.

Rabin (1993) described a model which captures this type of reciprocity. By using the 'psychological games' introduced by Geanakoplos et al. (1989),[7] Rabin developed a concept of equilibrium (the so–called *fairness equilibrium*) which reflects explicitly the notion that people desire to be kind to those who signal kindness through their actions and to hurt those who signal hostility through their actions. Although Rabin's model technically applies to 2–person games in the normal form, the intuition behind it can be extended to 3–person games by assuming that a person chooses not to cooperate (and hence to punish everybody) whenever he faces the alternatives of rewarding the partner who has cooperated or hurting the partner who has defected.

In the 3–person game under consideration, a similar assumption implies that if, after period μ, a persuasive player observes that one of his partners fails to reciprocate, then he will be punishing also the other by deciding to defect. The equivalent of relationships (5.16) and (5.17) can be, in this case, expressed as follows:

$$\forall t \in [1, \mu] \quad (\mu \geq 3): \quad g_{p,t} = 1 \quad \text{even if} \quad \forall k \neq p \quad g_{k,t-1} = 0; \tag{5.18}$$

$$\forall t \in [\mu + 1, T]: \quad g_{p,t} = \begin{cases} 0 & \text{if } \exists k \neq p \text{ s.t. } g_{k,t-1} = 0; \\ 1 & \text{otherwise;} \end{cases} \tag{5.19}$$

where μ must be not greater than the value given by (5.15).

The behavioural pattern resulting from (5.18) and (5.19) defines, clearly and completely, how (in a 3–person game with the linear utility function (5.14)) a strategy should evolve in order to be classified as persuasive.

[7] These games differ from the conventional ones because, in them, payoffs depend not only on players' actions but also on players' *beliefs*.

5.3 Game–theoretic consequences of the persuasive strategy

Purpose of this section is to show how a fully–rational (in the sense of orthodox game theory) player should behave against persuasive players, when there is complete information about the distribution of types in the population, and about their reciprocity–test–period if persuasive types.

Consider a N–person game with utility function (5.14):

$$U(g_{i,t}, y_t) = r(1 - g_{i,t}) + vy_t \quad \forall i \in N \qquad \forall t \in [1, T],$$

where $r > v$ and $Nv > r$.

I will prove that, in such a game, if the distribution of types is such that $(N-1)$ players are persuasive with *identical* or *different* reciprocity–test–period, and if this is common knowledge, then it will be fully–rational for the Nth player to cooperate for some periods.

Let us start such a demonstration by supposing that $N = 2$, and that player 1 is a persuasive player with reciprocity–test–period $\mu \leq T(2v-r)/v+1$.[8] Then, player 2 knows that his partner cooperates unconditionally (i.e., whatever player 2 does) for μ periods, and reciprocates thereafter (which means that from $\mu+1$ onwards, 1 plays whatever 2 has played in the previous period).

Given the (known) unconditional cooperation of player 1, it pays player 2 to defect until period $\mu - 1$. Indeed, if he does so, he gets $(\mu - 1)(r + v)$; if he does not (and, therefore, he cooperates), he gets $(\mu - 1)2v$; since $r > v$, the former is better (in terms of payoff–maximisation) than the latter.

But from period $\mu + 1$ onwards player 1's decision rule changes in the sense that $\forall t \in [\mu + 1, T]$: $g_{1,t} = g_{2,t-1}$. Knowing this, how should rational player 2 behave in the time interval between μ and T?

Since the last play is just a one–shot version of the game, a rational player should defect in the Tth period, no matter what the prior history might be. The latter along with the information about player 1's behaviour implies the following. If, from μ to $T - 1$, player 2 cooperates, he gets $2v$ throughout and $(r + v)$ in the last period. If, instead, from μ to $T - 1$, player 2 defects, he gets $(r + v)$ in period μ, and r thereafter. Thus, cooperation will be preferred to defection if: $[(T - 1) - \mu]2v + (r + v) > (r + v) + [T - (\mu + 1)]r$. Since $2v > r$, a fully–rational player should cooperate between period μ and period $T - 1$, and defect in T.

Keeping all the other aspects of the game constant, let us increase the

[8] This inequality is the equivalent of (5.15) with 2 rather than 3 players.

number of players to 3 (i.e., let us suppose $N = 3$). If, in this case, player 1 is the only persuasive player and both the other players are fully–rational, then the (complete–information) Nash equilibrium requires the latter to defect throughout the game.

However, if it is common knowledge that players 1 and 2 are both persuasive types with *identical* reciprocity–test–period μ (which, here, is not greater than the value given by (5.15)), then the fully–rational strategy is like that found before: rational player 3 should defect for $\mu - 1$ periods, cooperate from μ to $T - 1$, and defect in T. The rationality of the defective choice over the first $\mu - 1$ periods and in the final period derives from the fact that, by definition, $r + 2v > 3v$. The rationality of the cooperative choice between μ and $T - 1$ can be explained as follows. If, during this time interval, player 3 defects, he gets $(r + v) + [T - (\mu + 1)]r$; if, on the contrary, player 3 cooperates, he gets $[(T - 1) - \mu]3v + (r + v)$. Since $3v > r$, it is fully–rational to cooperate.

By generalising these results, we can state:

Proposition 5.1 *In a N–person game where each individual is characterised by the linear utility function (5.14), and where there is complete information about the distribution of types in the population, if $(N - 1)$ players are persuasive with identical (known) reciprocity–test–period μ, then full–rationality (based on maximisation of monetary payoffs) requires the Nth player to defect for $\mu - 1$ periods, to cooperate in the interval from μ to $T - 1$, and to defect in T.*

Let us suppose now that, *ceteris paribus*, the persuasive players' reciprocity–test–periods are different. Specifically, let us assume that the game involves three players, two of whom (players 1 and 2, for instance) are (known) persuasive types with $\mu_1 > \mu_2$.

In this case, player 3 knows that:
(a) until period μ_2 both the others cooperate, whatever he decides to do;
(b) from period μ_2 onwards, player 2 reciprocates the others' behaviour, in the sense that $\forall t \in [\mu_2, T]$: $g_{2,t} = \min\{g_{1,t-1}, g_{3,t-1}\}$, while player 1 continues to unconditionally cooperate up to period μ_1;
(c) from period μ_1 onwards, also player 1 reciprocates (in the sense of player 2) the others' behaviour.

Then, knowing (a), fully–rational player 3 should defect until period $\mu_2 - 1$. In fact, if he cooperates, he gets $(\mu_2 - 1)3v$; if he defects, he gets $(\mu_2 - 1)(r + 2v)$. Since $r > v$, the latter is preferred to the former.

Knowing (b), fully–rational player 3 puts forward his strategy in each period from μ_2 to $\mu_1 - 1$. If, in this time interval, he defects, he gets $r + 2v$ in μ_2 and

$r + v$ thereafter. If, instead, he cooperates, he gets $(\mu_1 - \mu_2)3v$. It can be seen that cooperation pays more than defection if $(\mu_1 - \mu_2)3v > r + 2v + [\mu_1 - (\mu_2 + 1)](r + v)$, which implies:

$$\mu_1 - \mu_2 > \frac{v}{2v - r}. \tag{5.20}$$

Thus, a fully–rational player cooperates from μ_2 to $\mu_1 - 1$ only if the number of periods between μ_2 and μ_1 is greater than the value given by (5.20).

Finally, knowing (c), fully–rational player 3 decides his strategy over the remaining $T - \mu_1$ periods. During this time interval, if he defects, he gets $(r + v)$ in period μ_1 and r thereafter. If, on the contrary, he cooperates, he gets $3v$ for $[(T-1)-\mu_1]$ periods and $(r+2v)$ in the final period (where, remember, he defects whatever the history of the game so far). Thus, cooperation is preferred to defection if $[(T-1)-\mu_1]3v + (r+2v) > (2v+r) + [(T-(\mu_1+1)]r$. Since $3v > r$, player 3 will maximise his pecuniary payoff if he cooperates from μ_1 to $(T-1)$ and defects in T.

The result found in case of three players can be generalised to a N–person game as follows.

Proposition 5.2 *In a N–person game where each individual is characterised by the linear utility function (5.14), and where there is complete information about the distribution of types in the population, if $(N-1)$ players are persuasive with different (known) reciprocity–test–period such that $\mu_1 < \mu_2 < ... < \mu_{N-1}$, then full–rationality requires the Nth player to behave as follows:*
α) to defect for the first $\mu - 1$ periods;
β) to cooperate in the interval included between two consecutive reciprocity–test–periods if the number of periods in each of these intervals is greater than the critical value given by (5.20);
γ) to cooperate from μ_{N-1} to $(T-1)$;
δ) to defect in T.

Thus, the (known) presence of persuasive players affects the behaviour of a fully–rational individual in that it pushes him to cooperate for some periods.

Despite the results obtained in this section, in the following discussion, whenever I refer to the Nash equilibrium, I will mean the equilibrium in the sense of orthodox game theory, i.e. the equilibrium without persuasive players (unless otherwise stated).

5.4 The persuaded player: a later–reciprocator of the others' maximum contribution

The persuasion hypothesis offers a specific characterisation of interactive behaviour, in the sense that the type of interaction in which the *persuasive* and the *persuaded* agents are engaged distinguishes my approach to cooperative behaviour from alternative ones. In the traditional models of altruism, for instance, the efficient cooperative outcome can be achieved if altruistic subjects face other people with their *same* attitude. Similar lines of argument apply to both the strategies hypothesis and the reciprocity theories, where only one type of individual needs to exist in order to justify cooperation in social dilemma situations.[9]

At odds with what happens in these theories, a persuasive process is meaningful and can succeed in promoting cooperation only if, in a group, besides a subject who, at the outset of the game, puts probability one on the fact that all the others, from μ onwards, will start contributing as much as he does until μ, there exist other members with 'selfish' attitudes but willing to change them. That is to say, in the relationship of persuasion, some group members must take the role of *respondents* in the sense that, during the interaction, they must be disposed to modify their own initial choices to make them converge on the same position as that of the persuader.

Under this aspect, the persuasion theory may appear similar to the imitation hypothesis. According to the latter, players can be distinguished in two categories: on the one hand, there are the 'good' players who get each period the most profitable outcome (and, hence, they do not change their strategy); on the other hand, there are the 'bad' players who modify their strategy by imitating the successful others.[10] It is exactly the interaction between 'good' and 'bad' players that, in the imitation story, provides reasons for the evolution of a particular strategy.

Apart from the requirement to have two distinct types of players, my hypothesis is quite different from the models of imitation. A first difference is that, while the process of imitation consists in a repetition of those strategies that have proved most successful, the persuaded subjects might also not imitate

[9]Particularly, in strategic contexts, the relevant figure is the player who places a positive probability on the fact that his partners are irrational and so he plays strategically against them. While, in reciprocity models, who matters is the individual willing to sacrifice his own material well–being to reward those who treat him adequately.

[10]See Offerman and Sonnemans (1995, p. 2) for a distinction between good players (the *ideal Bayesian observers*) and bad players (the *imperfect Bayesians*) in the imitation process.

successful others.[11] A further difference is that, while in the imitation models only one part of the interactive process is 'active' (in the sense that he does something) and this is the person who imitates his successful group members, in the persuasion process the position of both parts is active.

This section aims to define concisely and completely the category of the persuaded player.

Specifically, I shall assume that a persuaded player is an individual who selfishly–maximises his own payoff up to some period θ, and then he starts reciprocating his partners' *maximum* contribution.

Consider, for instance, the symmetric 3–person game with the linear utility function given in (5.14). Assume that one of the players is of the persuasive type and that he unconditionally cooperates for μ periods. Then, if a subject is observed to defect up to period $\theta \leq \mu$, and to cooperate thereafter or until at least one of his partners cooperates, then we can infer that the subject is a persuaded type.

Likewise, in a 3–person game where all individuals are characterised by the Stone–Geary utility function (5.8) and by identical initial endowments, if a subject is observed to contribute an amount correspondent to the Nash solution (5.9) for θ periods and to reciprocate the others' maximum contribution thereafter, then we can classify him as a persuaded player.

In general, in a repeated symmetric public goods game, the persuaded player's contribution, $g_{Pd,t}$, evolves as follows:

$$\forall t \in [1, \theta]: \qquad g_{Pd,t} = g^*; \tag{5.21}$$

$$\forall t \in [\theta + 1, T]: \qquad g_{Pd,t} = \max\{G_{k,t-1}\} \tag{5.22}$$

where g^* is the symmetric fully–rational Nash equilibrium of the game, and $G_{k,t-1}$ identifies the set of the contributions of the persuaded player's fellow members in period $t-1$.

But why a person should eventually modify his optimising behaviour and shift to this kind of reciprocal rule?

I argue that a subject may be induced to act according to the scheme just proposed for two reasons. The first one is that, since contributing is profitable only if it is chosen contemporaneously by other group members, the agent wants

[11] A persuasive player may get, in fact, the minimum possible payoff over a long sequence of periods, which depends on his reciprocity–test–period. It is therefore not on the basis of the persuasive player's former payoffs that, in my hypothesis, people decide to modify their strategy.

to be sure that some of his partners actually make this choice before abandoning the fully–rational strategy. Previous experimental studies suggest that people insist on entertaining doubts about the motives and character of their fellow members. In particular, the restart effect documented in public goods games indicates that some subjects are really interested in trying to reach the efficient outcome, and that it is their uncertainty about the others' behaviour that restrains them from contributing.[12] In the light of these observations, a persuaded player can be thought of as an individual whose uncertainty disappears in period θ.

The second reason that I suggest for explaining the persuaded agent's behaviour is simply that, at the start of the game, he has not grasped the benefits connected with contributions; being a fully–rational player, he has just chosen the strategy which maximises his monetary payoffs. Then, by observing one of his partners (the persuasive player) deviating constantly by such a strategy, he re–examines the incentives of the game, realises that contributing is mutually profitable, and so he starts following the other's behaviour.

Of course, it is not said that in a same group, besides a persuasive player, there exist persuaded players. The lack of these two different types of individuals leads the persuasive player to lower his contribution after period μ (when he is required to follow a reciprocal rule). The failure of attempts to persuade may explain the high initial levels of contributions and the decay phenomenon observed in a great number of repeated public goods experiments.

5.5 A comparison with other theories of cooperative behaviour

I shall now directly compare the theory of persuasion with some of the theories of cooperative behaviour considered in Chapter 4 (Section (4.2)). For each of these alternative theories, I will derive a number of testable restrictions by using, in each case, the non–linear utility function (5.8) as well as the linear function (5.14), so as to make it evident how they differ from my approach to cooperative behaviour.

5.5.1 Theories of commitment

The starting point of the persuasion hypothesis is that, for μ periods, an individual chooses to contribute a constant amount, irrespective of whether the others actually make this contribution (or, better, even if he believes that the

[12] Cf., Burlando and Hey (1997).

others will each play the fully–rational Nash equilibrium). For this reason, the persuasion theory is in the spirit of Kant's categorical imperative and may appear similar to commitment theories.

Nevertheless, while in Kant's moral and in theories of commitment "agents suppose that the others will continue to act as they do at the decision date",[13] in the theory of persuasion, a player must believe that, if he contributes a constant amount $\bar{g}_p$ up to period μ, all his partners will modify their initial 'selfish' choice and will contribute $\bar{g}_p$ from period μ until the end of the game. Only if his expectations are fulfilled, the persuasive player will keep on contributing $\bar{g}_p$ throughout the game.

Hence, the following proposition can be asserted:

Proposition 5.3 *Persuasion and commitment theories differ in that, under the former, a subject's unconditional contribution takes place only over μ periods and rests on his expectations that (from period μ forward) every other member of his group will converge towards his same contribution.*

The difference between my hypothesis and theories of commitment can be explicitly shown by considering which kinds of predictions the latter make in a symmetric 3–person game where each individual can be described by the Stone–Geary utility function given in (5.8).

The assumption upon which commitment theories are based (i.e., that an individual makes whatever contribution he would wish others to make, irrespective of whether they actually make this contribution) implies that, in order to decide how much to contribute, a Kantian type maximises utility function (5.8) subject to his belief that, $\forall k \neq i$ and $\forall t \in [1,T]$, $g_{k,t} = g_{i,t}$. By substituting this constraint into the objective function, the maximisation problem is equivalent to:

$$\max_{g_{i,t}} \quad (w - g_{i,t} - a)^{\alpha}(3g_{i,t} + b)^{1-\alpha}.$$

Differentiating with respect to $g_{i,t}$ and solving yields:

$$g^*_{c,t} = \frac{3(1-\alpha)(w-a) - \alpha b}{3} = \hat{g} \tag{5.23}$$

where $g^*_{c,t}$ is the optimal level of contribution under commitment theories and $\hat{g}$ is the symmetric social optimum described in Eq. (5.10).

Commitment theories predict, therefore, that an individual will contribute throughout the game a constant amount equal to the social optimum. This represents a testable restriction: if a subject is observed to contribute in any

[13] Cf., Laffont (1975, p. 431).

period an amount different from $\hat{g}$, then we can infer that he is not a Kantian type.

Furthermore, Eq. (5.23) implies that

$$\forall t \in [1,T]: \qquad \frac{\partial g^*_{c,t}}{\partial \sum_{k \neq c} g_{k,t-1}} = 0, \tag{5.24}$$

which is a further testable restriction: by regressing a subject's own contributions on those of his partners, if the slope coefficient differs from zero, this would mean that the subject is not following commitment theories' maximisation problem.

These predictions/restrictions (derived by using the particular utility function (5.8)) hold in any symmetric game, regardless of the function used to model each agent's preferences. Thus, for instance, if a Kantian type is involved in a finitely repeated 3–person game with the linear utility function (5.14), then he should cooperate throughout the game independently of the others' behaviour.

5.5.2 Theories of reciprocity

According to reciprocity theories, individuals should reward those who treat them adequately, and punish those who do harm them. Sugden's (1984) reciprocity principle (described in Chapter 4, Subsection (4.2.3)) affirms that anyone who benefits from the public good has *moral obligations* towards those who contribute; roughly speaking, this means, not that a person must always contribute, but that he must not take a free–ride when other people are contributing. Thus, under Sugden's theory of reciprocity the non–maximising/contributing behaviour would be followed because people share some principle of justice.

A similar argument suggests that the following proposition is true.

Proposition 5.4 *Persuasion and reciprocity theories differ because, under the former, an individual's contributions are not conceived as moral obligations towards the others, so that if everyone is free–riding such an obligation ends. Indeed, for μ periods, a persuasive agent is predicted to contribute even if all the other people in his group are free–riding.*

To see in which way the moral obligation upon which a reciprocator bases his choice affects his behaviour, consider once again utility function (5.8) and utility function (5.14). In either case, a subject who acts in accordance with Sugden's reciprocity principle maximises his own utility subject to the constraint: $g_{i,t} \equiv min(g^*_c, g_{k,t-1})$, where g^*_c is the optimal level of contribution under commitment theories, since it is (in Sugden's model) the contribution

that a subject would most prefer that everyone should make. Let $g^*_{r,t}$ be the solution of this maximisation problem.

Then, in the game with the non–linear utility function (5.8), we can prove that $g^* \leq g^*_{r,t} \leq \hat{g}$, where g^* and $\hat{g}$ are, respectively, the Nash solution (5.9) and the social optimum (5.10). In fact, if $min(g^*_c, g_{k,t-1}) = g^*_c$, agent i has the obligation to contribute at least g^*_c, which entails $g^*_{r,t} = \hat{g}$. If $min(g^*_c, g_{k,t-1}) = g_{k,t-1} = g^*$, agent i has the obligation to contribute at least the amount associated with the Nash solution. On the other hand, one can never have either $g^*_{r,t} < g^*$ (because the individual would find that self–interest dictates a larger contribution) or $g^*_{r,t} > \hat{g}$ (because the individual would be contributing more than he is obliged to). Hence, a subject who bases his choice on Sugden's reciprocity principle will contribute more than the amount of equilibrium only if everyone else in his group does so as well.

One interesting implication of the model is that an increase in the others' contributions should induce the reciprocator to increase his own contribution.[14] Hence, we can assume the following:

$$\forall t \in [1,T]: \qquad \frac{\partial g^*_{r,t}}{\partial \sum_{k \neq r} g_{k,t-1}} > 0, \tag{5.25}$$

which represents a testable restriction on reciprocal behaviour, contrasting with the analogous relationships (5.16) and (5.17) relative to the persuasion hypothesis, and with relationship (5.24) relative to commitment theories.

Suppose now that each individual's preferences can be described by the linear function (5.14). In case of 3 players, the moral constraint to which the reciprocator's utility is subjected leads $g^*_{r,t}$ to evolve as follows:

$$g^*_{r,1} = 1; \tag{5.26}$$

$$\forall t \in [2,T]: \quad g^*_{r,t} = \begin{cases} 0 & \text{if } \exists k \neq r \text{ s.t. } g_{k,t-1} = 0; \\ 1 & \text{otherwise.} \end{cases} \tag{5.27}$$

Thus, whenever a subject is observed not to cooperate in the first period and/or not to defect in any of the following periods as soon as he observes defection from one of his partners, this would imply that he is not a reciprocal type.

[14] See Sugden (1984) for a proof.

5.5.3 Theories of altruism

As said in Chapter 4 (Subsection (4.2.2)), altruism theories posit that individuals are motivated by a concern for other people's welfare. Hence, they maximise a utility function which is defined over a broader domain than that required by the traditional theory of pure self–interest, and which includes (in addition to their own) either the consumption or the utility of the others. This implies that, insofar as his contribution maximises the social welfare, an altruist may contribute to a public good even when he does not benefit directly from it.

This is not true for a persuasive player who decides how much to contribute by maximising *his own* pecuniary payoffs subject to an *a priori* set of beliefs about the others' behaviour throughout the game. Given these expectations, a persuasive player will contribute only if he (himself) derives benefits from such a decision.

Thus, we can affirm the following:

Proposition 5.5 *Persuasion and altruism theories differ in that, under the former, a subject having a well–defined (a priori) probability distribution over the others' type contributes only towards those public goods from which he (himself) derives private benefits.*

In a parallel to the analysis carried out with respect to commitment and reciprocity theories, let us identify the altruist's behaviour both in a 3–person game with the linear utility function (5.14), and in a 3–person game with the non–linear function (5.8).

In the former case, if we model altruistic preferences like in Eq. (4.3), and if the weight γ denoting the concern that the altruist expresses for the others' welfare is such that: $\gamma > (r-v)/(2v)$,[15] then an altruist's optimal decision, $g^*_{a,t}$, is to choose throughout the game the cooperative alternative independently of what the others do. Thus, provided that the altruist's weight on the others' welfare is greater than a critical value, we have:

$$\forall t \in [1,T]: \qquad g^*_{a,t} = 1, \tag{5.28}$$

which is a testable restriction for the game under consideration.

Let us now derive a testable restriction when the utility function has the non–linear form (5.8). With such a function, it seems reasonable to describe

[15] By using function (5.14)'s parameters values, Eq. (4.3) becomes:

$$V_i = r(1-g_{i,t}) + 3\gamma v y_t + (1-\gamma) v y_t + \gamma Pr_{k,t}.$$

Thus, cooperation is preferred to defection if and only if $3\gamma v + (1-\gamma)v > r$. The latter inequality is satisfied for each value of γ greater than $[r-v]/[2v]$.

altruistic preferences as below:

$$V(g_{i,t}, y_t) = U(g_{i,t}, y_t) \times [\prod_{k \neq i} U(g_{k,t}, y_t)]^{\gamma}. \tag{5.29}$$

Thus, the altruist's maximisation problem is:

$$\max_{g_{i,t}} \quad (w - g_i - a)^{\alpha}(y + b)^{1-\alpha} \times [\prod_{k \neq i}(w - g_k - a)^{\alpha}(y + b)^{1-\alpha}]^{\gamma}$$

or, equivalently

$$\max_{g_{i,t}} \quad \alpha \ln(w - g_{i,t} - a) + (1 - \alpha) \ln(y_t + b) + \\ + \gamma[\alpha \sum_{k \neq i} \ln(w - g_{k,t} - a) + (1 - \alpha) \sum_{k \neq i} \ln(y_t + b)].$$

If the altruist regards both the other members of his group in the same way (i.e., if both the others are the same for him), then the problem becomes:

$$\max_{g_{i,t}} \quad \alpha \ln(w - g_{i,t} - a) + (1 - \alpha)(1 + 2\gamma) \ln(y_t + b) + \gamma \alpha Pr_{k,t},$$

where $Pr_{k,t} \equiv \sum_{k \neq i} \ln(w - g_{k,t} - a)$ are the benefits that the other two group members derive from their private consumption.

The solution of this problem is:

$$g^*_{a,t} = \frac{(1 - \alpha)(w - a)[1 + 2\gamma] - \alpha(\sum_{k \neq i} g_{k,t} + b)}{\alpha + (1 - \alpha)[1 + 2\gamma]}. \tag{5.30}$$

It can be seen that if the game is symmetric (i.e., if $g_{j,t} = g^*_{a,t}$, $\forall j = 1, 2, 3$), then, for $\gamma = 1$, $g^*_{a,t} = [3(1 - \alpha)(w - a) - \alpha b]/3$ which is the social optimum, $\hat{g}$, given by (5.10). For values of γ between 0 and 1, the solution $g^*_{a,t}$ lies strictly between the Nash equilibrium given by (5.9) and $\hat{g}$.

The altruist's reaction function (5.30) allows us to derive a useful testable restriction. By differentiating such a function by the others' contribution, we find: $-\alpha/[\alpha(1 - \alpha)(1 + 2\gamma)]$. Since α and γ are both positive, this implies:

$$\frac{\partial g^*_{a,t}}{\partial \sum_{k \neq a} g_{k,t-1}} < 0. \tag{5.31}$$

Thus, by regressing a player's contributions over those of his partners, if the slope of the regression line does not come up to be negative, then we can infer that the player is not an altruist.

5.5.4 Reputation building/strategies hypothesis

Andreoni's (1988) strategies hypothesis is a theory of rational behaviour in the sense of Kreps et al. (1982). According to it, when the information about the other types is incomplete, if a fully–rational and selfish player believes that there is a small chance that the others are irrational (i.e., non–payoff maximising), it may be rational for him to contribute in order to build reputation. Toward the end of the game, however, the value of this strategic play decreases: in the last round, the free–riding strategy is always optimal.[16]

The following proposition can, therefore, be asserted.

Proposition 5.6 *Persuasion and strategies theories differ in that the former is not a theory of rational behaviour. A persuasive player's contributions are not caused by his uncertainty about the others' rationality. On the contrary, a persuader decides how much to contribute on the assumption that each of his partners is, at the outset of the game, fully–rational and as such he will play the dominant strategy for the first $\mu - 1$ periods.*

Let us consider how strategic play evolves in a 3–person game where each individual has the linear utility function (5.14). Both the hypothesis of incomplete information about the others' types and my own assumption that a strategic player needs to observe his partners defecting for two consecutive periods before dissolving his doubts about their rationality lead strategies to have a 3–phase structure. In an initial phase, which lasts two periods, strategic players cooperate in order to build reputation. In an end phase, which coincides with the last period, they defect, whatever the history of the game so far. The behaviour in the intermediate phase is determined by the contribution of the other players in the previous two periods, in the sense that, $\forall t \in [3, T-1]$, a strategic player cooperates only if one of his fellow members has not defected in each of the previous two periods. Hence, if $g^*_{s,t}$ indicates the contribution of a strategic player in period t, we have:

$$\forall t = 1, 2: \qquad g^*_{s,t} = 1. \tag{5.32}$$

$$\forall t \in [3, T-1]: \quad g^*_{s,t} = \begin{cases} 0 & \text{if } \exists k \neq s \text{ s.t. } g_{k,t-2} = g_{k,t-1} = 0; \\ 1 & \text{otherwise.} \end{cases} \tag{5.33}$$

[16] Because there is no future, contributing to the public good cannot induce any further contribution by the other players, and so there is no reason for a strategic player to contribute.

$$g^*_{s,T} = 0. \tag{5.34}$$

Thus, unless a subject is observed to cooperate for the first two periods, to defect in the last period, and to defect in any period from 3 to $(T-1)$ as soon as he observes one of his partners to defect for two consecutive periods, then the subject cannot be considered a strategic type.

Likewise, in a 3–person game where all individuals have the non–linear utility (5.8) and identical initial endowments, we can think of strategies as characterised by a 3–phase structure. In this case, however, given the non–binary dimension of the strategy set, it is not easy to define clearly and completely how strategic behaviour should evolve in the intermediate phase.[17] No problems exist, instead, for the initial phase and the end phase.

Indeed, in the end phase (which coincides with the last period of interaction), a strategic individual must play the fully–rational Nash equilibrium contributing g^* to the public good (where g^* is given by (5.9)), whatever the history of the game so far.

In the initial phase (which lasts two periods), a strategic player, who believes that both his partners are irrational and that as such they will contribute an amount greater than the dominant–strategy equilibrium, tries to build reputation as a cooperative type himself. If he supposes that the other players will each contribute the amount $\bar{g} > g^*$, then (in order to build reputation) he must contribute $\bar{g}$ as well. This implies the following optimisation:

$$\max_{g_{i,t}} (w - g_{i,t} - a)^{\alpha} (y_t + b)^{1-\alpha}$$

subject to:

$$\sum_{k \neq i} g_{k,t} = 2[\delta \bar{g} + (1-\delta) g^*],$$

and

$$g_{i,t} = \bar{g},$$

where δ $(0 < \delta \leq 1)$ is the probability that i attaches to the likelihood that his fellow members which each play $\bar{g}$ rather than g^*.

By substituting the constraints into the objective function, the maximisation

[17] I will come back to this point later on.

problem becomes:

$$\max_{\bar{g}} \quad \alpha \ln(w - \bar{g} - a) + (1 - \alpha) \ln(\bar{g} + 2[\delta \bar{g} + (1 - \delta) g^*] + b).$$

The solution to this is:

$$g^*_{s,t} = (1 - \alpha)(w - a) - \frac{2\alpha(1 - \delta)}{1 + 2\delta} g^* - \frac{\alpha b}{1 + 2\delta}. \tag{5.35}$$

It can be seen that if $\delta = 1$, then $g^*_{s,t} = [3(1 - \alpha)(w - a) - \alpha b]/3$ which (in this symmetric case) corresponds to the unique social optimum, $\hat{g}$, given by (5.10). For values of δ between 0 and 1, the solution $g^*_{s,t}$ lies strictly between the Nash solution, g^*, given by (5.9) and $\hat{g}$.

By differentiating (5.35) with respect to δ, we find a positive derivative. This confirms that higher values of δ increase the amount of contribution.

Insofar as we know the value of δ, Eq. (5.35) represents a testable restriction: given δ, if an agent is observed to contribute, in the first two periods, an amount different than stated in (5.35), then we can infer that he is not following strategic optimisation.

As for the evolution of strategies in the intermediate phase (which lasts from period 3 to period $T - 1$), this depends on the others' behaviour, in the sense that whenever a strategic player observes that, for two consecutive periods,[18] one of his fellow members does not contribute as much as him, then he changes his decision.

The theory does not specify the form of this change. Kreps et al. (1982) confine their attention to games of the prisoner's dilemma type. In such games, a player has to choose only from two alternatives, which implies that, if he wants to stop cooperation, then (by excluding randomised strategies) he can just defect. On the contrary, in the game under consideration, if a player intends to change his decision from one period to the next, then he can choose among a range of possible contributions and he can follow several alternative decision rules.

With respect to this issue, previous experimental work helps. Keser (1999), for instance, found that behaviour in this intermediate phase can be described by reciprocity in the sense that it changes in the direction of the others' average contribution in the previous period.

[18] The length of this time interval derives from my own assumption

5.6 Final thoughts

The economic analysis of cooperative behaviour is still an open matter. In the past few years various theories have been proposed to explain such behaviour but none of them appears able to provide reasons for all, or even most, of the observed regularities.

The hypothesis of persuasive behaviour suggested in this chapter aims to unravel previously inexplicable cooperation. It is relatively simple, and rests on two principles that seem to have a strong common–sense. People understand which group's action is most beneficial for them and, consequently, they try to induce their partners to perform it.

Since in a repeated game where verbal communication is forbidden, the only way that a player has got to signal information to the others is by his choices of moves as the game proceeds, a persuasive agent contributes a constant amount up to a certain period μ, by expecting that, from such a period onwards, the other players will each be willing to modify their initial 'selfish' choice and to contribute as much as he does. Then, if his expectation is fulfilled, the persuasive agent perseveres with his constant contribution as long as he observes that everyone else does so as well. Otherwise, he modifies his contribution in the direction of the others' average contribution in the previous period.

Hence, the persuasion theory is compatible with both the observed successes and the observed failures of voluntary cooperation.

In principle, it deserves to be taken seriously as one among several ways to explain individuals' underlying motivations for choosing actions that do not maximise their monetary payoffs (especially in early stages of repeated social dilemma games).

In practice, since it generates testable predictions/restrictions which differ from those of earlier theories, its capability of accounting for previously inexplicable behaviour can be empirically verified.

The next two chapters report on a series of public goods experiments which are designed in order to test directly the persuasion hypothesis and separate it from earlier theories of cooperative behaviour.

Such an empirical analysis will start by considering a very simple 3–person public goods game where each player is characterised by the linear utility function (5.14).

Then, the voluntary contribution mechanism environments in which people interact will be complicated by taking into account games with the Stone–Geary function (5.8).

Chapter 6

My 'simple' experiment

6.1 Introduction

This chapter aims to verify by means of the experimental method if subjects' attempts to persuade can explain why, in early decision rounds of repeated public good games, people are systematically found to contribute more than predicted by complete–information non–cooperative models. Or, to put this in a slightly different way, the purpose of this chapter is to examine whether the provision of public goods through voluntary contributions might be motivated by an individual's expectations of inducing his partners (who are free–riding) to do what is best for the group (and, hence, *for themselves*) rather than selfishly maximising own earnings.

As a first step towards more decisive conclusions about the possibility of persuasion being an explanation for previously inexplicable cooperation, the domain for such a test is, here, a simple three–player dilemma game with the linear utility function given in (5.14).

There are some novelties in the way in which the experiment is designed that make it appropriate for my research question. Indeed, it combines behaviour with belief–elicitation and (what is less common) with an investigation into subjects' motivations for their own decisions. Asking subjects to motivate their decisions is not only a very straightforward way to obtain insights into their decision rules, but it should also induce them to be more concentrated on the game, and to think seriously about the problem that they face.

Key aspects of the experimental implementation of the voluntary contribution mechanism used in this study are reported in the next section, where I describe the constituent game (around which the experiment was constructed), and I explain the organisation of each single decision round.

6.2 The constituent game and the decision round's structure

The laboratory version of the voluntary contribution mechanism utilised in the experiment presented here was implemented in a sequence of 6 repeated games, or 6 *phases*.

Each phase consisted of 10 repetitions of a simple 3–person symmetric game, where the subjects' decisions were binary and a representative individual's payoff function in any one period was given by the linear function (5.14):

$$U(g_{i,t}, y_t) = r(1 - g_{i,t}) + vy_t.$$

The values of the parameters v and r were chosen to meet the inequalities $r > v$ and $3v > r$. Thus, each subject had a dominant strategy to contribute zero (which, using backward induction, is also the unique subgame–perfect equilibrium of the repeated game). The maximum payoff to the entire group is attained, however, if in each repetition all subjects contributed one.

The decision of having groups of size three rather than two (and, therefore, of making subjects playing a public goods game rather than just a prisoner's dilemma) finds its justification in the fact that the former allow for the verification of some issues that cannot be explored using the latter. With 3–person games, for instance, we can verify how cooperative subject i reacts when only one of his fellow members cooperates. Does subject i reciprocate the cooperative partner? Or, instead, does the defecting behaviour of the other partner lead him to defect and to punish, in such a way, also the other group member? Since the answer to this question is crucial in testing some of the predictions/restrictions identified in Chapter 5, the choice to have groups with more than two subjects seemed to be unavoidable.

Then, among all possible group sizes, I decided for groups of three in order to create a simple and clear decisional environment.[1] Almost all previous public goods experiments are based on N–person games (where N is usually equal to 4 or 5), in which subjects can provide the public good at any level between their initial endowment and zero. In such settings, each player must think about the behaviour of many other people whose contribution decisions lie over a wide range. So, how the individual player interprets (and reacts to) the behaviour of his partners is difficult to observe in these experiments. Since the study reported in this chapter is only a first test of the persuasive behaviour hypothesis, the

[1] "Simplicity is a good feature of experiments. You are more likely to understand what you have learned" (Ledyard 1995, p.176).

experiment was designed so as to implement the simplest possible environment. The latter appears to be a game where, every period, subjects have to choose only from two alternatives while keeping their mind on the decision processes of only other two people, which is exactly the 3 (players) × 2 (strategies) game used here. This (simple) starting point represents a necessary step towards more complicated stages. The experimental investigation of any new hypothesis should proceed gradually. The inefficiency of doing everything in one single step is, in fact, well known.

Despite its simplicity, the possibility that, in my constituent game, a subject chooses to cooperate out of *confusion* cannot be discarded. This choice is, indeed, the only 'mistake' which—from a game theoretic point of view—a subject can make. Hence, one may interpret an individual's cooperative choice as an error and attribute it to his confusion. In addition, because of its repeated nature, even egotistical subjects may decide to cooperate in (early periods of) the game because they play strategically and want to conceal their rationality. Thus, alternative equilibrium concepts, like the sequential equilibrium/strategies hypothesis, may be invoked to explain cooperation.

A very straightforward way to obtain insights into subjects' decision rules, and better understand the relationship between people's actions in public goods settings is to directly ask them both to motivate their own decision and to speculate about the motivation of the others. For this reason, I included in the experiment a questionnaire through which I collected motivational information. In addition, I elicited expectations about the others' future behaviour.

Thus, any single period was divided into two stages: the *decision* stage and the *questionnaire* stage.

The decision stage

Instead of telling the participants the payoff function (5.14), I presented the game in terms of Fig. 6.1, where X denotes the private good, Y the public one, and where the binary choice between X and Y is equivalent to the choice between contributing either nothing ($g_i = 0$) or all ($g_i = 1$) to the public good.[2]

Hence, in each decision stage, each subject was asked to choose between X and Y, by knowing that if he chose X, only he would receive r tokens;[3] while if he chose Y, each subject in his group would get v tokens, with $r > v$ and

[2]Actually the game was shown in Fig. 6.1's format only in the instructions. Particular values for v and r were inserted, in fact, during the experiment. I will say more about the values of these parameters later.

[3]The token was the unit of experimental money. The exchange rate between token and real money was: 1 token = 0.2 pence

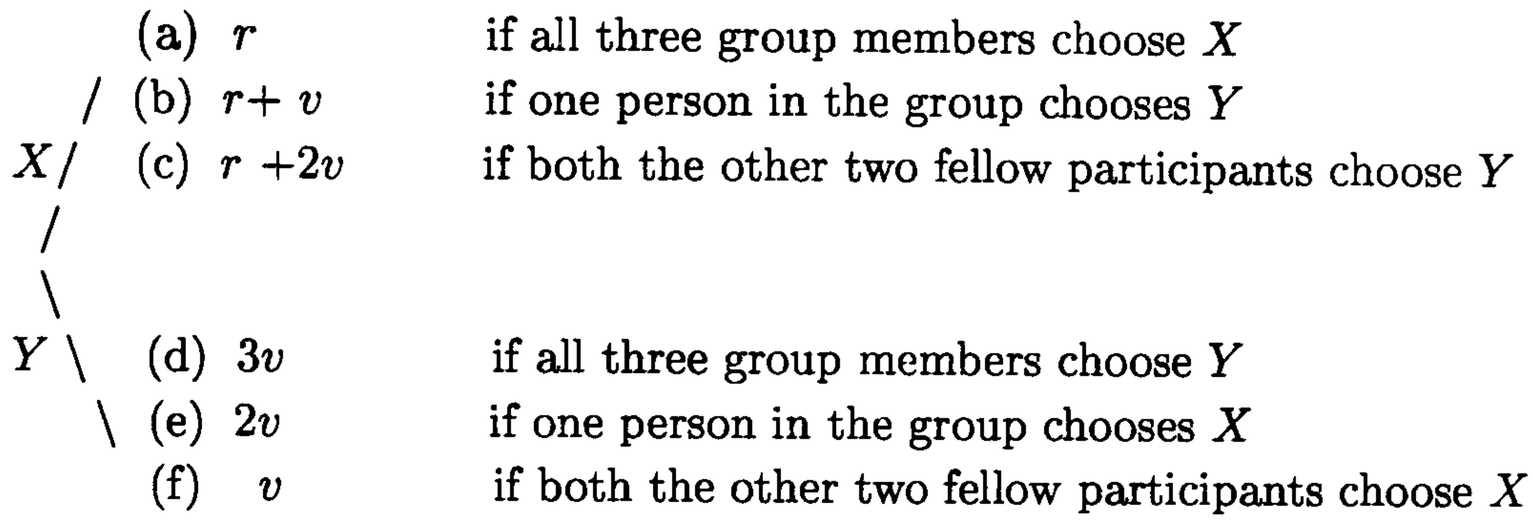


Figure 6.1: Form in which the constituent game was presented to subjects.

$3v > r$. Given these inequalities, the fully–rational Nash equilibrium (generated by own–payoff maximisation) is for each player to select X, while the social optimum requires each subject to choose Y.

The questionnaire stage

Every period—after having selected an action from the two that were permitted, and having observed the decisions of his partners—each player was requested:
1) to indicate the main reason for his own period–decision;
2) to guess the motivation behind his fellow members' decision;
3) to predict his fellow members' next decision.

Subjects answered the first two questions by picking out—for each of them—one of five different alternatives (see Tables A.1 and A.2 in Appendix A). They had to type X or Y on their keyboard to state their expectations about the others' next choices.

In order to avoid the criticism of not providing participants with (monetary) incentives to answer the questionnaire seriously and honestly, subjects were compensated both for right guesses of the others' motivations and for accurate predictions. In particular, participants were told that, each round, in addition to their earnings from the decision stage, they would be rewarded with 30 tokens if they guessed correctly the motivations lying behind the last choice of both their partners, and that they would get a further 30 tokens if their predictions about both the others' next choice turned out to be right.

Such an incentive scheme could be accomplished (and subjects themselves could believe in it) only if the others' possible motivations were in a limited number. For this reason, rather than allowing participants to write down the motivations for their decisions on their own, I gave them a list of alternatives which is perfectly symmetric to the list given for answering the second question, as can be seen by comparing Tables A.1 and A.2.

A look at these tables reveals that they lack any answer explicitly corre-

sponding to the persuasion hypothesis (e.g., "I want to persuade the others"), or to any other of the hypotheses of cooperative behaviour presented in the previous chapter. I deliberately omitted these kinds of motivations from the questionnaire in order to avoid the criticism of influencing the subjects' thinking process by suggesting them specific behavioural rules.

While the examination of the relationship between a subject's action and his beliefs about the motivations underlying others' behaviour is peculiar to my experimental design, there have been previous experiments which attempted to investigate the relationship between an individual's decision and his expectations about his fellow members' next decision.[4] However, in a departure from most of these earlier studies, in the experiment presented here, expectations were elicited in period t for period $t+1$ *after* a subject had observed the actual choices made by the others in t.[5] Such a procedure was followed for practical reasons: that is, in order to make subjects answer all three questions in a unique stage so as to divide each period into two parts only. Since the question asking the subject to speculate about the others' motivation can be answered only *after* observing the others' decision, and since it seems reasonable to ask people to motivate their own decision once they have already taken it, if I want people to answer all three questions in one single step, then the questionnaire stage must inevitably come after the decision stage. If, by following Croson (1998a, 1998b), I had elicited expectations before decisions, each single period would have been composed of three parts: a first part in which subjects had to predict the others' next choice; a second part in which they had to take their own decision; and, finally, a third part in which they had to answer the other two questions. This would have made the experiment itself too complicated for the subjects, and the program (for running it) too difficult to implement.

The only problem with such a procedure is that we do not get information about expectations prior to the first decision–round.

---

[4]For economic examples of experiments which elicited expectations in public goods settings see Messick et al. (1983); Schroeder et al. (1983); Poppe and Utens (1986); Fleishman (1988); Weimann (1994); and Croson (1998; 1998b). There is also a large literature in psychology that concentrated on the elicitation of individuals' expectations in prisoner's and social dilemma games (cf. Kelley and Stahelski (1970); Kuhlman and Wimberley (1976); Dawes et al. (1977); and Messé and Sivacek (1979)).

[5]In both Croson's experiments (1998a; 1998b) the guess treatment consisted of an additional estimation stage before each game.

6.3 Experimental parameters and procedures

The computerised experiment was run in the laboratory of the Centre for Experimental Economics (EXEC) at the University of York (UK).[6] It was organised in two sessions with 12 subjects for each session (so that a total of 24 people took part in this experiment). Each session consisted of two perfectly identical subsessions.

6.3.1 About the first subsession

The first subsession was divided in three different phases (or supergames), each of which was a 10–fold repetition of the game reported in Fig. 6.1.[7]

Phases were distinguished according to:
1) the groups composition; i.e., subjects interacted in the same group during an entire phase, while groups were randomly formed anew from a phase to the next one;
2) the value of the mrs, which (given utility function (5.14)) is simply the ratio v/r.

The main reason for having Partners treatments is that, in non cooperative games, a player can pursue a persuasive strategy only if his fellow members do not change after each period. Participants were informed that they would remain in the same group throughout each phase. They were also told that the composition of their groups would randomly change from a phase to the next one, and that they could not expect to meet the same fellow members again in a later phase.[8]

As for the mrs, this was varied by changing, from phase to phase, the reward v for choosing the cooperative alternative, which took values of 25 in the first phase, 33 in the second phase, and 44 in the third phase. The return r from selecting the defecting alternative was instead always 50, both across and during the phases.[9]

---

[6]The program for running the experiment was developed by myself and by the EXEC programmer Norman Spivey.

[7]Given the simplicity of the problem presented to the subjects and the provided definition of persuasive play, ten iterations are considered enough to understand if a subject follows a persuasive strategy and if he succeeds in it.

[8]Given the experimental design (3 supergames played sequentially by groups of three players), subjects never would meet the same partners in a later phase only if at least 12 people participate in the experiment. This is exactly the number of subjects who took part in each session.

[9]The values of v were chosen so that the inequalities $r > v$ and $Nv > r$ hold. Since $r = 50$ and $N = 3$, the lower and upper bounds between which v has to lie are 17 and 49, respectively. Furthermore, into this range, I decided for the described values on the basis of a pilot experiment.

The difference in the *mrs* allows me to investigate if subjects modify their pattern of behaviour according to phase and, in particular, if they become more cooperative from a phase to the next one. Let $P_{i,t}(j)$ be player i's penalty for choosing the cooperative alternative Y in round t of phase j when both his two fellow members choose X; that is, $P_{i,t}(j) = (v - r)$. Then: $P_{i,t}(1) = -25$, $P_{i,t}(2) = -17$, $P_{i,t}(3) = -10$. Therefore, while in the first phase a cooperative player can lose 25 tokens each round, in the last phase his loss is 10 tokens. Hence I expect to see a greater number of participants deciding to contribute in the last phase. This opinion is reinforced by the fact that incentives to defect (given by $-P_{i,t}(j)$)[10] decrease, whilst gains to be achieved by coordinating on the Pareto optimal outcome increase across phases.

Subjects were shown the game they had to play only when the new phase started. Such a procedure was followed to eliminate the possibility of 'order effects' due to participants' knowledge of increasing values of v.

6.3.2 About the second subsession

By definition, somebody cheats if in the last period he plays the dominant strategy. A crucial difference between strategies hypothesis and the other theories of cooperative behaviour discussed in Chapter 5 is that, whilst according to the former a player should always cheat, the latter never ask for cheating if all group members are cooperating.

In order to verify if (and how) the behaviour of a cooperative subject would change whenever he does not cheat but his partners do, and study, therefore, whether the attitudes of 'experienced' subjects towards choosing the cooperative alternative is influenced by previous interactions, I included in my experiment a second subsession which was an *exact* repetition of the first one[11] and *crucially* I did not inform participants of this. Indeed, any information about the contents of subsession 2 was given after the end of subsession 1. At the start of the experiment, however, subjects were told that the experiment would be longer than that just explained in the instructions, and that they would be required to stay in the laboratory for an hour more.[12]

Such an experimental design permits us to verify:

---

[10] Player i's incentive to defect is given by his payoffs when he plays the dominant strategy while the other members of his group cooperate, i.e. $r + 2v$, minus his payoffs if he continues playing cooperatively, i.e. $3v$. Therefore we have: $[(r + 2v) - (3v)] = (r - v) = -P_{i,t}(j)$.

[11] This means that the same three phases described in Subsection (6.3.1) in the same succession, with the same parameter values, and with the same group composition were performed twice.

[12] This bit of information was given in order to avoid deceptive practices that are not recommended in economics experiments (cf., Hey (1991, p. 119); and Ledyard (1995)).

1) if subjects defect in the last period of each phase of the first subsession (the presumed end of the plays) even if the partners are cooperating;
2) in what ways cheating on the part of his fellow members influences the behaviour of each player in the second identical subsession. If a player does not cheat, but his partners do, he might take a dislike to the others and so modify his previous attitudes towards them. This change will be reflected in his successive choices.

6.3.3 Subject pool

Subjects were undergraduate and graduate students of the University of York. Students were volunteers recruited by announcements on several information boards in the University buildings and by mail–shot invitations. Upon their arrival at the laboratory of EXEC, they received a copy of the instructions for the first subsession.[13] These instructions were also read aloud. The understanding of these instructions was checked in a short computerized exercise program. Each subject had to go through this program individually. Then, the subjects were randomly arranged in four groups of three, and played the first 10–period game (i.e., the first phase). They made their decisions anonymously via computer terminals and did not communicate with each other in any other way.

During the experiment, at the end of each round, participants were shown information (by means of a 'results table' displayed on their computer screen) on their own experimental earnings for the round just finished as well as on the choices and corresponding earnings of each of their partners.

It was explained that all the decisions took in each round (i.e., the choice between X and Y as well as the answers to the questionnaire) were binding and that end–of–experiment rewards would be based on the sum of earnings from all rounds. Subjects were told that the value of each experimental money (i.e. of each token) was 0.2 pence. The average payoff, earned in about two hours, was approximately 10 English pounds.

6.4 Constructing the alternative hypotheses

The linear utility function around which this experiment is constructed has been used in Chapter 5 in order to illustrate alternative hypotheses of cooperative behaviour. In that context, I have also derived, for each theory, a

[13] Complete copy of the instructions is reported in Appendix A.

number of testable predictions/restrictions. The latter, together with the motivational/beliefs data collected through the questionnaire, provide the set of criteria to be used here in order to construct the different hypotheses, or *player types*, open to investigation.

Viewing each hypothesis as a player type should lead us to consider it invariant, for any individual subject, over the six phases; it seems, in fact, reasonable to expect a player to retain his own type throughout the experiment.

Some features of the design might, nevertheless, affect such a reasonable expectation. First, the fact that phases were characterised by different mrs might induce subjects to modify their behavioural pattern according to phase; from the initial laboratory work on free–riding, it is well known that the size of the mrs is one of the most significant factors in influencing people behaviour.[14] Second, to the extent that phases 4, 5, and 6 were a new start of phases 1, 2, and 3 respectively, this might have a bearing on the behaviour of a player, who (in the light of his experience) might be induced to change his strategy (and, therefore, his type) when he re–faces the same partners. Third, previous interaction, although with different partners, might influence a subject by leading him to re–consider his strategy when a new phase starts. Thus, if one of these three circumstances takes place, we might well observe a player to modify his type throughout the experiment.

6.4.1 The persuasive type

In this experiment, a persuasive strategy is concisely and completely defined by restriction (5.15) (on the maximum number of periods for which a persuasive player can cooperate unreciprocated, i.e., without being followed by each of his partners), and by the behavioural pattern resulting from (5.18) and (5.19).

Substituting the parameters T, r and v with their actual values into (5.15), we obtain $\mu \leq 6$ in phases 1 and 4, $\mu \leq 8$ in phases 2 and 5, and $\mu \leq 10$ in phases 3 and 6.[15]

Thus, in each phase of this experiment, a player will be taken to be of the persuasive type if he is observed:

(**a**) to choose the cooperative alternative in each period from 1 to μ regardless of the others' behaviour throughout;

(**b**) to shift to defection in any of the following periods as soon as he observes defection from one of his partners. Unless both the others are cooperating, this

[14] On this topic see, for instance, Ledyard (1995).

[15] Since μ can be an integer number only, the decimal numbers have been rounded to the smallest nearest integer.

shift must be detected no later than the 6th round in phases 1 and 4, and no later than the 8th round in phases 2 and 5. This means that if a subject is found to cooperate, unreciprocated, for more than 6 periods in phases 1 and 4, and for more than 8 periods in phases 2 and 5, then he cannot be classified as a persuasive type. As for phases 3 and 6, in them, a persuasive player might cooperate, unreciprocated, for all 10 periods.

The questionnaire's data allow us to add further testable restrictions on persuasive behaviour.

Let us first consider the relationship between choices and expectations. My definition of a persuasive player as an 'hard–wired' individual who is allowed to have some kind of discretion/rationality only at the outset of the game implies that, once the player has solved his first period's maximisation problem and has decided his strategy (on the basis of an *a priori* and well–defined set of beliefs about the others' behaviour throughout the game), he cannot anymore optimise. His *a priori* beliefs may well change when the game is actually played, but this change has no bearing on his 'automatic' behaviour. The predefined set of instructions given to the persuasive player require him to cooperate for μ periods and to reciprocate the others' behaviour thereafter. Such a program does not allow for changes in behaviour based on changed expectations.

One of the most relevant consequences of this definition of a persuasive strategy is that it induces a player to never defect and *to never expect defection* as long as he observes cooperation from each of the other players. The persuasive agent has, in fact, solved his maximisation problem on the assumption that both he and his partners will play the cooperative alternative from period μ onwards; therefore, if, during this time interval, the others are observed to cooperate, the persuasive player can neither deviate from cooperation, nor can expect them to defect.

As for the reason indicated for his period–decision, in order to be classified as persuasive a subject must motivate his cooperation of the first μ periods by selecting either alternative C[16] or alternative E,[17] providing, in the latter case, a personal motivation somehow related to the persuasion hypothesis (e.g., "I wanted to persuade the others", or "If we all choose Y, then each will benefit"). The cooperative choices (if any) made after μ can be explained by C, E as well as by A.[18] While the (possible) defective choices must be motivated by A, B,[19]

[16]That is: "If all 3 group members took this choice, we would obtain the highest payments".

[17]This alternative gave the subject the possibility to type his own reason on his keyboard, if none of those suggested could justify his period–decision.

[18]By choosing A, a subject attributes his period–decision to the past behaviour of his fellow members.

[19]That is: "Whatever my fellow–participants would have chosen, this choice assured me

or E.

6.4.2 The reciprocal type

The behaviour of a subject who bases his choice on Sugden's reciprocity principle evolves as given by (5.26) and (5.27). The latter represent two testable restrictions which are used here in order to identify a reciprocal type. Thus, in each phase of this experiment, a subject will be classified as a reciprocator if he is found:
(**a**) to cooperate in the first round;
(**b**) to shift to defection in any of the following rounds as soon as he observes defection from one of his partners.

Also under reciprocity theories, a subject can change his behaviour (and shift, therefore, from the initial cooperation towards defection) only on the basis of the others' *observed* behaviour. The moral obligation that the reciprocator has towards his partners leads him to cooperate until at least one of the others does so as well.

As for his motivational answers, a reciprocator must match his cooperative choices with alternative A or C, and his defecting choices (if any) with alternative A, B, or D.[20] Either choice can be justified by alternative E; in this case it must be possible to relate the provided personal reason with reciprocal behaviour (e.g., "I have just reciprocated the choice of my partner", or "I wanted to punish (to reward) my partners who were hostile (kind) towards me").

6.4.3 The strategic type

As pointed out in Chapter 5, strategic behaviour is characterised by a 3–phase structure. In the setting under consideration, these 3 phases are concisely and completely defined by (5.32), (5.33) and (5.34). Thus, in each 10–period phase of this experiment, a subject will be taken to be of the strategic type if he is observed:
(**a**) to cooperate for the first two periods whatever his partners do, but only if he expects cooperation from each of them; the latter implies that unless a subject, in $t = 1$, expects both his partners to cooperate in $t = 2$, then the subject cannot be considered of the strategic type;[21]
(**b**) to defect in the last round, whatever the history of the game so far;

higher payments than those I could obtain by choosing the other action".

[20] That is: "This choice gave my fellow–participants the lowest payments".

[21] A strategic individual attaches a small probability to the likelihood that his playing partners are not fully–rational (i.e., they use dominated strategies). This implies that he must expect cooperation from them.

(**c**) to defect in any period from 3 to 9 as soon as he observes one of his partners to defect for two consecutive periods.

As for his motivational data, a strategic type has to motivate his cooperative choice of the first two periods by picking out alternative C, and his defecting choice of the last period by picking out alternative B. The choices in the intermediate time interval must be matched with alternative A or C, if cooperative choices; and with alternative A or B, if defecting choices. In any period, any choice can be justified by E, insofar as the reason given on his own is related with strategies (e.g., "I want the others to think that I will cooperate if they do").

6.4.4 The Kantian type

In this experiment, a Kantian type must cooperate throughout each phase, independently of (his expectations about) the others' behaviour.

He must choose alternative C (or E) to explain his constant cooperation.

6.4.5 The altruistic type

The evolution of altruistic behaviour in the game used here is given by (5.28). According to the latter, if γ (i.e., the altruist's weight on the others' welfare) is such that $\gamma > [r - v]/[2v]$, then an altruist's optimal decision is to choose the cooperative alternative throughout each phase independently of what (he expects) the others (to) do.

By using this experiment's parameters' values, we obtain: $\gamma > 0.5$ in phases 1 and 4; $\gamma > 0.25$ in phases 2 and 5; and $\gamma > 0.125$ in phases 3 and 6.

Since γ is an unknown variable, I cannot really detect an altruistic type in this experiment.[22] The only possible test on altruistic behaviour is the following: unless a subject cooperates for all 10 periods of a phase regardless of (his expectations about) the others' behaviour, then the subject cannot be an altruist.

6.4.6 The persuaded type

Restrictions (5.21) and (5.22) of previous Chapter 5 concern a persuaded player's behaviour. Applied to this experiment, they lead us to consider of the persuaded type a subject who, in any 10–period phase, is observed:
(**a**) to defect until some period θ, regardless of (his expectations about) the

[22] A subject might, indeed, be an altruist and, nevertheless, decide to defect throughout a phase because his own γ is less than the critical value corresponding to that phase.

others' behaviour;
(b) to cooperate thereafter if at least one of his partners is observed to cooperate.

A persuaded type must pick out alternative B to justify his prior defection, and alternative A or C to justify his later cooperation.

6.4.7 The Nash type

In Chapter 5 (Subsection (5.3)), I have shown that, in a N–person game, where there is complete information about the distribution of types in the population, if $(N-1)$ players are persuasive, then it will be fully–rational for the Nth player to cooperate for some periods.

In this experiment, the lack of complete information about the others' types leads a subject who wants to selfishly maximise his own–payoff (whatever the decision of the others) to defect in each repetition. Thus, a subject will be considered of the Nash type if he defects throughout a phase whatever (his expectations about) the others' behaviour.

As for his motivational answers, a Nash type player must match his repeated defection with alternative A, B, or E (giving, in the latter case, a personal motivation related to his desire to maximise his own monetary payoff).

6.4.8 Summary of the criteria used in identifying a player type

By summarising, three aspects of the data are here taken into account in order to classify a subject as a particular type:

(1) the maximum periods of *unreciprocated cooperation* that a subject exhibits;

(2) the relationship between a subject's choice and his partners' observed behaviour;

(3) the reason a subject indicates for his own period–decision; i.e. the way in which he answers the first question of the questionnaire stage.

As for point (1), three of the theories considered allow us to derive a testable restriction relative to this issue. Such a restriction can be expressed, in general form, as follows: "a subject cannot be considered of type φ if he cooperates, unreciprocated (i.e., without being followed by each of his partners), for more periods than τ." Then, if φ stands for *persuasive*, τ equals 6 in phases 1 and 4; 8 in phases 2 and 5; and 10 in phases 3 and 6. If φ stands for *reciprocal*, τ equals 1, regardless of the phase played. If φ stands for *strategic*, τ equals 2, whatever the phase played.

This restriction does not apply to a Kantian type or to an altruistic type, who, on the contrary, must be observed to cooperate *even if unreciprocated* throughout each phase. For the altruist, such an unconditional cooperation takes place only if his weight on the others' welfare is greater than a critical value.

As for point (2), four of the six cooperative hypotheses investigated here predict a specific kind of relationship between choices and others' behaviour. A persuasive type and a reciprocal type must not defect as long as both the other group members do not defect. A strategic type must not defect as long as one of his partners does not defect for two consecutive periods. A persuaded type, who defects in all periods from 1 to θ, must not defect afterwards if at least one of his partners cooperates.

As for point (3), the match motivation–decision that must be observed for each type has been given in the previous subsection.

6.5 Aggregate results

In presenting my results, I shall start with an aggregate description of the data. Figure 6.2 shows the aggregate average of cooperative subjects separately for the 10 periods of the 6 phases.

An analysis of these data makes the following observations possible.

Observation 6.1 *In all periods, the percentage of people contributing is significantly different from zero, but also significantly less than the efficient outcome of 'all group cooperating'.*

Observation 6.2 *Cooperation tends to fall with repetition in general, although a closer look at the period–by–period data reveals that, during this downward trend, the percentage of cooperators remains at an (almost) constant level for more than two consecutive periods.*

Observation 6.3 *Subjects do not play their dominant strategy in the final period; in particular, a substantial amount of contributions is reported at the end of subsession 1 (the 'presumed' end of the game).*

Observation 6.4 *Each positive variation in the mrs is accompanied by an increase in cooperation.*

It can be shown that the increase in cooperation across phases is statistically significant. Pooling data over phases, (under the assumption of independent observations) I use a test of proportions, which is based on a binomial distribution

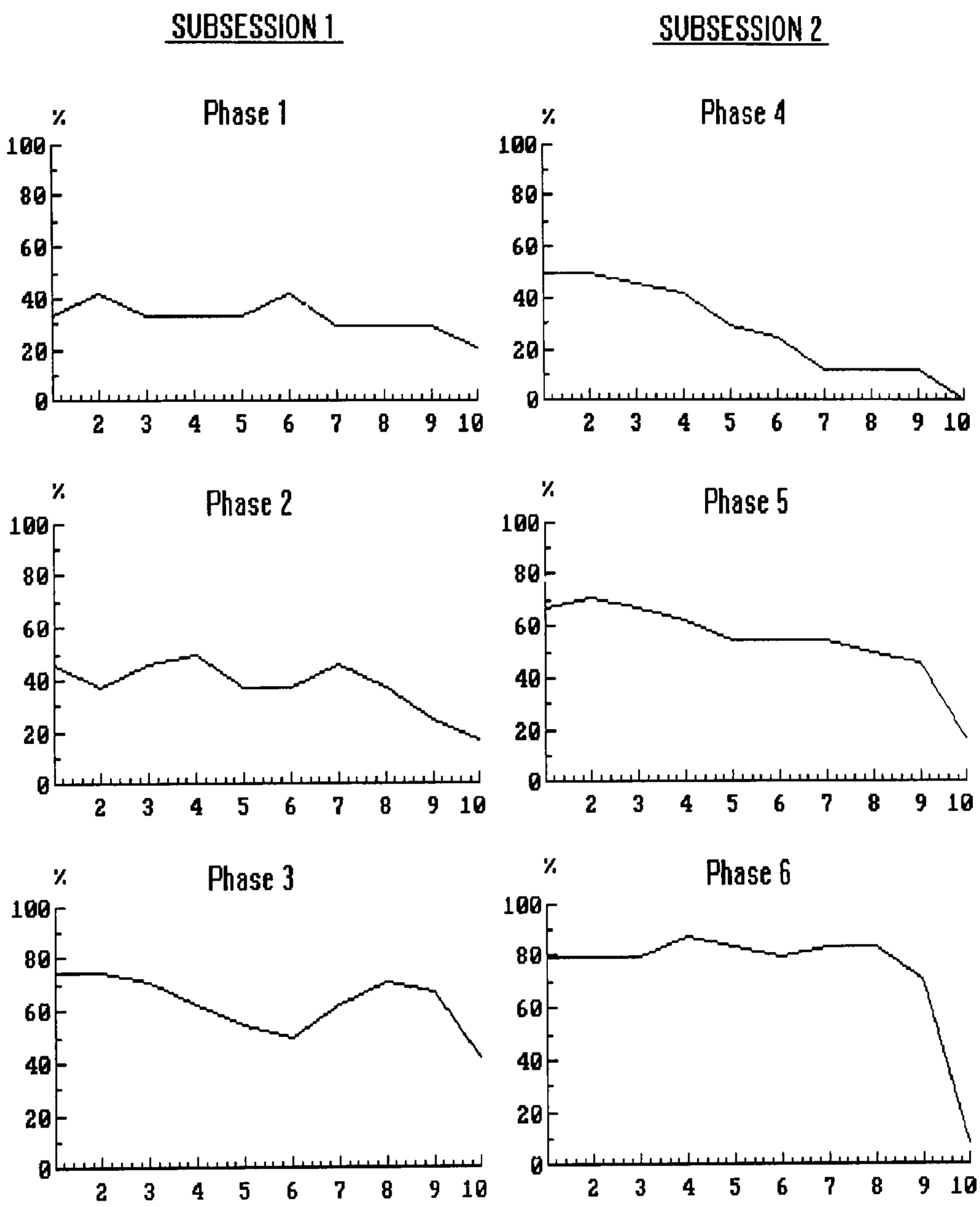


Figure 6.2: Average of subjects cooperating in each phase.

(calling giving a success and free–riding a failure), to compare cooperation in each pair of phases.

Let p_i and p_j denote the population proportions of successes (i.e., of cooperators) in phases i and j (respectively), where the population proportions of successes in phase i ($\forall i = 1, 2, ..., 6$) is defined as the number of cooperative choices detected in phase i out of a total of 240 choices (10 period–choices $\times$ 24 subjects) divided by 240. I test the null hypothesis $H_0 : p_i - p_j = 0$ against the alternative $H_1 : p_i - p_j > 0$. The decision rule is to reject H_0 in favour of H_1 if

$$\frac{\hat{p}_i - \hat{p}_j}{\sqrt{\hat{p}_0(1-\hat{p}_0)(\frac{n_i+n_j}{n_i n_j})}} = \overline{z}_{i,j} > z_\alpha,$$

where $\hat{p}_i$ and $\hat{p}_j$ are the observed sample proportions for phase i and phase j, respectively; $n_i = n_j = 240$ is the number of observations; $\hat{p}_0$ is the estimate of the common proportion under the null hypothesis, given by:

$$\hat{p}_0 = \frac{n_i\hat{p}_i + n_j\hat{p}_j}{n_i + n_j} = \frac{240\hat{p}_i + 240\hat{p}_j}{480},$$

and $\overline{z}_{i,j}$ is the value obtained by comparing the proportions of cooperators in phase i and phase j.

Results of these tests are depicted in Table 6.1.

Table 6.1: z–statistics for the differences in cooperators' proportions between phases.

Subsession 1	*Subsession 2*
$\overline{z}_{1,2} = 1.53^{+}$	$\overline{z}_{5,4} = 6.02^{***}$
$\overline{z}_{3,2} = 5.39^{***}$	$\overline{z}_{6,5} = 4.28^{***}$
$\overline{z}_{3,1} = 6.86^{***}$	$\overline{z}_{6,4} = 9.86^{***}$

$^{***}\ p < 0.001$ $^{+}\ p < 0.10$

As can be seen, the null hypothesis of equality of cooperators' proportions between phases can be rejected at significance levels beyond the 1% for all comparisons, except for the one between phase 1 and phase 2, where the null hypothesis is rejected only at the 10% level.

All observations from 6.1 to 6.4 are in line with the findings of previous experimental studies. What I want to test here is whether such a behaviour can be due to the fact that cooperators, after having understood the benefits of the efficient strategy, continue cooperating in order to induce defectors to do likewise.

The finding that the percentage of cooperators versus defectors does not drop immediately to zero as well as the result that there is a substantial amount of cooperation in the tenth period of each phase of the first subsession might somehow support my hypothesis if two conditions were verified. First, if the constancy of the cooperators' percentage across more periods of the same phase means that they are actually the same subjects who keep on taking the cooperative choice. Second, if the last round's cooperators are people who were cooperating during most previous periods.

These (and other) questions can be answered by turning to the analysis of behaviour across individuals.

6.6 Individual results

It is well documented in the literature that individuals differ consistently with respect to behaviour in these kinds of experiments. Generally speaking, a first relevant distinction can be made between *individualistic* and *cooperative* subjects. Then, within the latter category, a further distinction regards individuals' motivations for cooperating.[23] In this subsection, I shall investigate whether any of the rival hypotheses presented in Section (6.4) can explain this experiment's data and provide justification for the observed cooperative behaviour.

I first discuss the relationship between an individual's choice and his responses to two of the questions in the questionnaire stage. These answers will be used to verify subjects' understanding of the game. Next I shall present a classification of subjects according to the criteria listed in Subsection (6.4.8). After that, an analysis of behaviour across subsessions will allow for further conclusions.

6.6.1 Testing subjects' confusion

In my experiment, the only 'mistake' a subject can make—from a game theoretic point of view—is to choose the cooperative alternative Y. Therefore, one may interpret an individual's choice of Y as an error and attribute it to his confusion.[24] First, I must rule out the possibility of confusion being the explanation of the observed cooperation and then I can go on to investigate which theory of cooperative behaviour best justifies subjects' decisions.

Participants' motivations of their own period–decisions as well as their guesses about others' motivations allow for a test of the *confusion hypothesis.*

A subject's reason for his own period–decision shall be defined as 'inconsistent' if it cannot explain in any way the decision for which it is given.

Let us consider the different alternatives among which participants are asked to pick out in order to motivate their own choices, and that are reported on Table A.1 of Appendix A. Inconsistency is found, for instance, when a subject motivates his choice of Y by selecting alternative B which is clearly referring to the Nash equilibrium strategy; or when he declares to choose an action because of the past behaviour of his fellow members who have never played that action.

[23] The claim that individuals differ with respect to their motivation for cooperating coincides with everyday experiences as well as with the results of experiments made, among others, by Marwell and Ames (1981); Carter and Irons (1991); Andreoni and Miller (1993); Brandts and Schram (1997). On this topic, see also Weimann (1994); Offerman et al. (1996); and Burlando and Hey (1997).

[24] See Andreoni (1995a); Palfrey and Prisbey (1996, 1997); and Brandts and Schram (1997) for a discussion about 'confusion' and 'noise' in public goods experiments.

The choice of alternative E when the specification of the other reasons is missing is also considered inconsistent.

Analysis of this kind of inconsistency reveals that 34 out of a total of 1440 motivations are inconsistent, and that only 11 of these concern the cooperative choice. Experimental data, then, show that:

- in the first phase, only 1 out of 24 subjects selects *twice* Y by giving each time inconsistent motivations for this choice;
- in the second phase, another subject—who exhibits a cooperative behaviour over all 10 periods—provides *5 times* inconsistent motivations;
- the remaining *4 inconsistencies* are spread over all six phases and across different individuals.

These findings lead to the following observation.

Observation 6.5 *People in my experiment do not seem to be cooperative out of their confusion about the monetary incentives of the game.*

To test the robustness of such an observation, I shall use subjects' responses to the second question of the questionnaire stage; i.e., their guesses of the others' motivations. Table A.2 in Appendix A lists all alternatives available to subjects' in order to answer this question.

In this case, I shall adopt the expression *inconsistent guess* to indicate a contradiction between the others' observed behaviour and the alternative picked out in order to motivate that behaviour.

Results from this investigation demonstrate that, throughout all the experiment, only 2.7 percent of subjects who choose Y provide inconsistent guesses of their counterparts' motivations. Among them, there are both the subjects who answered inconsistently the first question of the questionnaire. Such a result leads me to attribute their cooperative choice to a mistake and to consider them as *confused* subjects (see the classification of subjects reported in Table 6.2).

Of particular interest here is the percentage of cooperators who motivate the others' defecting choice by picking out alternative B from those listed in Table A.2. Such a percentage refers, in fact, to people who play cooperatively even if they have grasped the true incentives of the game. Their choice of Y cannot be therefore attributed to any mistake. My data reveal that, throughout all experiment, 57 percent of cooperating people make this combination.

An important motivation for eliciting this kind of guesses is to see whether or not people, forced to think about the motivations behind others' behaviour, modify their own behavioural pattern.[25]

[25]For instance, cooperating people might understand the equilibrium of the game *after*

If the simple act of asking for reasons behind others' choice makes individual's behaviour really different, I should observe an *instantaneous* shift from an action to the other soon after the cited match is done. However, this instantaneous shift is rarely observed. Particularly, no immediate shift occurs in phases 1 and 3; whilst two shifts are observed in phase 2 (one from Y to X, the other from X to Y).

A last result relative to this study deserves to be underlined. A subject is generally found to believe that others think like him when they make the same choice. This observation is in line with the findings of previous experiments.

The analysis carried out in this subsection allows me to distinguish between *selfish*, *confused* and *cooperative* subjects. The purpose of the next subsections will be to explore further the source of the observed cooperation.

6.6.2 Testing the alternative hypotheses of cooperative behaviour

Here (unless otherwise stated) I use the criteria reported in Subsection (6.4.8) to classify subjects. In some cases, such criteria do not permit an explanation of the data and, therefore, do not allow the determination of which of the rival hypotheses best tracks the individual's behavioural pattern. When this occurs, the subject will not be arranged in any type/category, but will be considered *unclassifiable.*

In addition, it is impossible to impute a specific type to a subject who cooperates in all 10 periods of a phase, while he observes and expects cooperation from each of his partners: except the strategic hypothesis, all the other cooperative hypotheses investigated in this study can account for this kind of behaviour. For this reason, whenever I observe groups whose members cooperate and expect cooperation one another throughout a phase, I will not arrange any individual subject in a particular category, but all members of the group will be considered *always–cooperators.* An always–cooperator identifies, therefore, a subject who (at odds with an unclassifiable individual, whose behaviour cannot be explained by any theory) satisfies the criteria used here to identify an altruist, a Kantian player, a reciprocator and a persuasive individual, so that he may be any of these types.

---

having matched other's choice of X with alternative B. Similarly, defecting people might appreciate the benefits of cooperation *after* the symmetrical (and opposite) match: other's choice of Y—(because of) alternative C.

First subsession's results

I shall begin my test of the various behavioural hypotheses by taking into account data from the first subsession. Table 6.2 shows the distribution of the subjects among the different hypotheses/types in each of the three phases of subsession 1.[26]

Table 6.2: Distribution of subjects across the three phases of subsession 1.

	Phase 1	Phase 2	Phase 3
Nash	8	7	5
Confused	1	1	0
Reciprocal	0	1	0
Kantian/Altruist	0	2	1
Strategic	2	1	3
Persuasive	3	3	4
Persuaded	0	3	0
Unclassified	8	6	5
Always–cooperators	2	0	6

First of all, it can be seen that (in agreement with the data on average cooperation) the number of subjects playing the dominant–strategy Nash equilibrium decreases across phases. This confirms that people are less 'selfish' when the penalty for choosing the cooperative alternative decreases.

As far as the cooperative subjects are concerned, inspection of Table 6.2 reveals the following.

a) Two subjects (one in phase 1, another in phase 2) are confused. As emphasised in the previous subsection, such an attribution derives from the fact that these subjects provide repeatedly inconsistent reasons for their own choice as well as inconsistent guesses of their counterparts' motivations.

b) Only one subject (and just in one of the three phases) behaves in such a way that the classification criteria employed here lead me to consider him of the reciprocal type.

c) Few subjects are of the Kantian or of the altruistic type. These subjects are found both to cooperate throughout a phase even if, in some periods, they expect defection from their partners and to motivate their constant cooperative

[26] Remember that the three phases of the first subsession were distinguished according to the group composition and the value of the mrs.

choice by picking out, in each period, alternative C. In this case, it is impossible to distinguish whether such subjects cooperate because they follow a Kantian morality or because they have altruistic preferences.

d) In all 3 phases, we observe subjects classifiable as strategic individuals as well as subjects of the persuasive type.

e) Three subjects (all concentrated in phase 2) behave as the persuaded type.

f) Some subjects (whose number is particularly high in phase 3) may be of the altruistic, Kantian, reciprocal or persuasive type: they cooperate throughout a phase while observing and expecting cooperation from each of the others. As such, they are considered always–cooperators. Notice that the detection of this behavioural pattern implies that some of the participants in this experiment do not cheat in the final period of a phase, but (against any possible rational model) they prefer to cooperate.

g) Finally, a remarkable number of participants do not present readily interpretable data and, consequently, are considered unclassifiable. Among these unclassifiable patterns of behaviour, we observe some homogeneity which is worth emphasising. For instance, a few subjects are found to start cooperating at a later stage of a phase, and to keep on this choice, unreciprocated, for some consecutive periods. All these subjects motivate their later shift to cooperation on their own and, in all cases, they provide a personal reason related to their desire to signal a willingness to cooperate.[27]

In summary, Table 6.2 shows that subjects are heterogeneous with respect to their motivations for cooperating, and that persuasion is a strategy followed by some of the people classified as cooperators. In addition, (in line with the corresponding value of μ) the highest tendency to persuade is observed in phase 3: in this phase, a subject can pursue a persuasive strategy—and cooperate unreciprocated—for all 10 periods.

The distribution of types across phases reported in Table 6.2 further suggests that there is heterogeneity not only across individuals but also, within the same individual, across phases. Thus, although it seems reasonable to expect a player to retain his own type over phases, most subjects in this experiment appear to change their type according to phase. This can clearly be seen from Table 6.3, which displays the evolution of each subject's behaviour over all three phases of subsession 1, and also reports, for each subject, the period, ν, (if any) in which he switches from cooperation to defection.

[27] One can find examples of this kind of behaviour in Appendix C (where I report all individual data collected from this experiment) by looking at the data of Subject 15 in phase 1,

Table 6.3: Classification of each subject according to his type across the three phases of subsession 1.[a]

Subject	Phase 1	Phase 2	Phase 3
1	P ($\nu = 5$)	P ($\nu = 5$)	P ($\nu = 4$)
2	U	R ($\nu = 2$)	N
3	U	U	S ($\nu = 3$)
4	N	U	AC
5	S ($\nu = 3$)	N	AC
6	N	N	N
7	N	C	U
8	N	U	P ($\nu = 9$)
9	U	U	U
10	P ($\nu = 5$)	U	U
11	U	N	N
12	P ($\nu = 7$)	U	AC
13	C	N	U
14	AC	P ($\nu = 9$)	N
15	U	Pd	U
16	N	N	P ($\nu = 10$)
17	S ($\nu = 10$)	S ($\nu = 3$)	P ($\nu = 10$)
18	U	P ($\nu = 5$)	AC
19	U	K/A	AC
20	U	Pd	S ($\nu = 10$)
21	N	Pd	AC
22	AC	K/A	S ($\nu = 10$)
23	N	N	K/A
24	N	N	N

[a] At each letter corresponds a type:
C=Confused; K/A=Kantian/Altruist;
N=Nash; P=Persuasive; Pd=Persuaded;
R=Reciprocal; S=Strategic; U=Unclassified;
AC=Always-cooperators.

Examination of Table 6.3 reveals that only three subjects retain their own

Subject 8 in phase 2, and Subject 7 in phase 3.

type throughout subsession 1. All the other subjects are found to fall, in each phase, under a different category/type. As pointed out on p. 123, two characteristics of the design might plausibly explain such a finding: the difference in the *mrs* across phases, and the experience gained by the subjects as the experiment proceeds.

Second subsession's results

There are two ways of analysing the data relative to subsession 2. One is to consider the two subsessions of my experiment completely independent so as to identify each player type as described in Section (6.4). By taking the same definition of type as reference point, we can examine if the experience gained by a subject in the first subsession affects his thinking process and his attitude towards the others so as to induce him to *modify* his type when he (unexpectedly) re–faces the same partners and the same parameter values.

The other method is to see phases 4, 5 and 6 as continuations of phases 1, 2 and 3 respectively, in which case some player types are predicted to behave in the former phases differently than in the latter phases.[28] By looking at the data of the second subsession from this perspective, we can verify if subjects actually modify their type, or if, instead, their behavioural change (if any) is explicable in terms of the different prediction of the theory.

I shall analyse subsession 2's data by using both procedures and I will show that they generate almost the same classification of the subjects.

Let us start by assuming complete independence between subsessions. Using the criteria listed in Subsection (6.4.8) in order to classify subjects, we obtain Table 6.4 and Table 6.5. Table 6.4 shows the distribution of the subjects among the different hypotheses/types in each of the 3 phases of subsession 2. Table 6.5 displays the evolution of each subject's behaviour from the fourth to the sixth phase.

A comparison between Table 6.2 and Table 6.4 reveals that, while the distribution of players of the Nash–type in any phase of subsession 2 mirrors pretty much that observed in the corresponding phase of subsession 1, this does not apply to the distribution of cooperative types. In phases 4, 5 and 6, subjects are found to distribute themselves among the alternative cooperative hypotheses in a quite different way than that observed in phases 1, 2 and 3 respectively.

The main difference is detected between phase 3 and phase 6. In the latter phase, there is a significant decrease in the number of always–cooperators as

[28] I will specify later what each theory prescribes in a phase of the second subsession, in the light of the player's experience in the corresponding phase of the first subsession.

Table 6.4: Distribution of subjects across the three phases of subsession 2.

	Phase 4	Phase 5	Phase 6
Nash	9	7	3
Confused	0	0	0
Reciprocal	0	0	0
Kantian/Altruist	0	3	1
Strategic	4	6	14
Persuasive	5	3	1
Persuaded	0	0	2
Unclassified	6	4	3
Always–cooperators	0	1	0

well as in that of persuasive players in comparison with the former phase. At the same time, the number of players of the strategic type is noticeably higher in phase 6 than in phase 3 (14 versus 3).

A comparison between Table 6.3 and Table 6.5 (i.e., a study of the same subject's behaviour across phases) reveals that all 6 always–cooperators of phase 3 become strategic players (with $\nu = 10$) in phase 6; all of them were cheated by one of their partners in the last period of phase 3. The same change towards strategies can be observed for 2 subjects who were persuasive, but did not succeed in their attempts of persuasion, in phase 3. Investigation into the motivations that these subjects provided for their own decisions demonstrates that all of them shift to defection in the final period of phase 6 because of the past behaviour of their fellow members. They matched, indeed, their choice of X with alternative A.

Thus, a first important result emerging from the analysis carried out in this subsection concerns the great effect of previous experience on people's behaviour: subjects are found to remember what happened in the first part of the experiment and, in the second part, they change their behaviour and their attitude towards the others' interests.

Such a result can be interpreted in terms of the learning models discussed in Chapter 4 (Subsection (4.2.4)), and re–stated as follows: as subjects in my experiment gain knowledge about the others' type, the path describing their behaviour converges towards the non–cooperative, fully–rational Nash equilibrium.

Table 6.5: Classification of each subject according to his type across the three phases of subsession 2.

Subject	Phase 4	Phase 5	Phase 6
1	S ($\nu = 10$)	U	S ($\nu = 10$)
2	N	K/A	S ($\nu = 10$)
3	U	P ($\nu = 9$)	S ($\nu = 10$)
4	S ($\nu = 10$)	P ($\nu = 5$)	S ($\nu = 10$)
5	S ($\nu = 3$)	N	S ($\nu = 10$)
6	N	N	N
7	N	U	N
8	U	U	S ($\nu = 10$)
9	S ($\nu = 10$)	S ($\nu = 10$)	N
10	P ($\nu = 5$)	S ($\nu = 10$)	Pd
11	N	U	S ($\nu = 10$)
12	U	AC	S ($\nu = 10$)
13	P ($\nu = 5$)	K/A	S ($\nu = 10$)
14	U	N	S ($\nu = 10$)
15	N	S ($\nu = 10$)	U
16	N	N	K/A
17	P ($\nu = 6$)	S ($\nu = 10$)	Pd
18	N	P ($\nu = 5$)	S ($\nu = 10$)
19	U	K/A	S ($\nu = 10$)
20	U	N	U
21	P ($\nu = 4$)	N	S ($\nu = 10$)
22	P ($\nu = 5$)	S ($\nu = 10$)	P ($\nu = 6$)
23	N	S ($\nu = 10$)	U
24	N	N	S ($\nu = 10$)

Notice that the learning occurring here is only learning about the others' behaviour, not also learning about the incentives of the game and the dominant–strategy equilibrium. Indeed, Observation 6.5 as well as the analysis of inconsistent guesses carried out in the previous subsection make it evident that individuals in my experiment are not confused about the game. The learning going on here is, therefore, not of the kind described by Andreoni (1988, 1995a) or by

those learning models which require confusion to be the principal explanation for cooperation.[29]

On the other hand, a comparison of Tables 6.3 and 6.5 also reveals that some subjects become more cooperative in the second subsession. Consider, for instance, subjects 4 and 21: both are Nash type in phase 1, while they are classified (respectively) as strategic type and persuasive type in phase 4. Observe, also, subjects 13 and 23: both are Nash type in phase 2, and become (respectively) a Kantian/altruist type and a strategic type in phase 4. Reason for these changes can be found in the *restart effect*. As emphasised in Chapter 4 (Subsection (4.3.1)), a break between two subsessions allows for what is called in psychology 'cognitive dissonance': a moment in which players stop the continuity of decisions and/or actions and re–think about what to do.

To show that the restart effect plays actually a role in my experiment, let us compare the number of people choosing the cooperative alternative at the end of each phase of subsession 1 with the number of people choosing the same alternative at the outset of the corresponding phase of subsession 2. Inspection of Figure 6.2 clearly suggests that the average cooperation in period 10 of phases 1, 2 and 3 is less than that in period 1 of phases 4, 5 and 6 respectively. This is confirmed by Table 6.6, which reports the number of cooperators in the last period of each of the first three phases and the number of cooperators in the first period of each of the second three phases.

Table 6.6: The 'restart effect' in this experiment.

	Phase 1	Phase 2	Phase 3
No. of cooperators in $t = 10$	4	4	9
	Phase 4	Phase 5	Phase 6
No. of cooperators in $t = 1$	12	16	19

Besides providing justification for the more cooperative behaviour exhibited by some subjects in the second subsession than in the first one, the presence of such a restart effect supports the observation that learning in the sense of Andreoni is not at work in this experiment. If this were the case, the subjects should be unaffected by the restart and they should continue to play the free–riding equilibrium. But this is not what we observe.

If (rather than considering the two subsessions independent) we see phases 4, 5 and 6 as continuations of phases 1, 2 and 3 respectively, this may have a

[29] Cf., e.g., Palfrey and Prisbey (1996, 1997).

bearing on the way in which we can interpret the data relative to subsession 2. Having observed that experience affects behaviour, since some cooperative types are predicted to behave differently when they are 'experienced' individuals than when they have no experience at all, we can verify whether the change in behaviour observed for most subjects can be 'captured' and explained in terms of these different theoretical predictions.[30]

Let us therefore consider what each hypothesis prescribes with respect to behaviour in the second subsession, in the light of the players' experience in the first subsession.

Two out of the six cooperative types investigated in this study (i.e., the Kantian type and the altruistic type) are predicted not to modify their pattern of choices in subsession 2, regardless of the knowledge acquired in subsession 1. The Kantian type (being guided by Kant's categorical imperative) must keep on undertaking that action which, when generalised, yields the best outcome to all agents. The altruistic type (having the others' welfare inherent in his utility function) should not be affected by the (possible) observed lack of other altruistic group members: he is expected to cooperate also throughout a subsession 2's phase, insofar as his weight on the others' welfare remains greater than $[r - v]/[2v]$.

In contrast with these two types, the behaviour of each of the other cooperative types changes with experience. In case of reciprocal, strategic and persuaded players, experience modifies behaviour at the outset of a phase, in the sense that these types of players may start a phase of subsession 2 with a choice which does not correspond to the first choice assigned to them by the classification criteria discussed in Section (6.4). According to these criteria, in the first period of any supergame, a reciprocator and a strategic individual must cooperate, while a persuaded agent must defect. To what extent experience changes these predictions is discussed below.

As far as a reciprocal player is concerned, we know that the moral obligation which he has towards the others leads him to cooperate whenever both the others cooperate too; on the other side, such an obligation ends as soon as he observes defection from one of them. This implies that if a reciprocator observes one of his fellow members to defect throughout (or in late periods of) any of the first three phases, then he can use this bit of information in the corresponding second subsession's phase and start such a phase with defection. Thus, a subject can be observed to defect in the first period of any phase of subsession 2 and be (nevertheless) considered of the reciprocal type if, in the

[30] For instance, a subject classified as a Nash type in Table 6.4 might be a strategic type whose partners have defected in the first ten periods.

corresponding phase of subsession 1, he was a reciprocator whose partners have not cooperated.

A strategic player, on the other hand, cooperates, in early repetitions of a game, only if he attaches some small probability to the likelihood that each of his playing partners is 'irrational'. This implies that if, in any of the first three phases, the strategic subject learns that at least one of the others is not of the irrational type, then, at the outset of the corresponding second subsession's phase, he has no longer interest in cooperating so as to build up reputation. Thus, a defecting choice by a strategic player in the first period of phase 4 (say) is admitted insofar as, in phase 1, he followed, unsuccessfully, a strategic behaviour.

Similarly to the reciprocator and the strategic player, the choice of the persuaded agent at the outset of any phase of subsession 2 depends on the path of behaviour that he has observed in the corresponding phase of subsession 1. A persuaded player should start a phase of subsession 2 cooperating (rather than defecting), if at least one of his playing partners has ended up the corresponding phase of subsession 1 with cooperation.

As far as the persuasion hypothesis is concerned, we know that one of the most relevant consequences of the program defining a persuasive strategy is that it induces a player to never defect and *to never expect defection* as long as he observes cooperation from each of the other players. A prediction of defection made when both the others are cooperating contradicts the assumptions upon which the persuasive player builds his optimisation problem of the first period. As such, it cannot be justified. Relying on this definition of a persuasive strategy, a subject cannot be of the persuasive type if he is found to conclude a phase of the first subsession with defection, while both he and his partners were cooperating up to then.

Experience gained in the first subsession modifies the latter restriction in that, in the light of this experience, the persuasive player can modify the set of expectations about the others' behaviour which underlie the maximisation problem he faces at the outset of a phase in subsession 2 (when he is allowed to have some discretion/rationality).

If his playing partners have defected throughout (or in later periods of) a phase of subsession 1, the persuasive player may well expect them to defect in the final period of the corresponding subsession 2's phase. As a consequence, his decision problem changes. He must now choose either to cooperate or to defect on the assumption that his partners will each defect until $(\mu - 1)$ and they will continue to do so unless he cooperates throughout, in which case they

will cooperate from periods μ to $(T-1)$ and defect in T, when he defects as well. Thus, it will pay him to cooperate (rather than defect) if and only if:

$$(\mu-1)v+[(T-1)-(\mu-1)]3v+r \geq Tr,$$

which yields:

$$\mu \leq (T-1)\frac{3v-r}{2v}+1. \tag{6.1}$$

The latter equation represents a testable restriction on persuasive behaviour with experienced players: given the defecting behaviour of his group members in a phase of subsession 1, if a player expects them to defect in the last period of the corresponding subsession 2's phase (whatever the history of the game so far), then, in order to be classified as a persuasive type, the player must be observed not to cooperate unreciprocated for more periods than those given by (6.1).

Substituting the parameters T, r and v with their actual values into (6.1), we obtain $\mu \leq 5$ in phase 4, $\mu \leq 7$ in phase 5, and $\mu \leq 9$ in phase 6.

The instructions that the 'hard–wired' persuasive individual must now follow during the game are: cooperate until μ; reciprocate the others' behaviour from $(\mu+1)$ to $(T-1)$ (which means: defect in any period in this time interval as soon as you observe defection from one of your partners); defect in T, whatever the others' behaviour so far.

Thus, a subject who defects in the last period of any phase of subsession 2, *despite his own and his partners' cooperative behaviour until then*, can be considered of the persuasive type if:
(a) he is found, in $t=9$, to expect both his partners to defect in the last period;
(b) such an expectation is justifiable by the defecting behaviour actually shown by his group members in the corresponding phase of subsession 1;
(c) in providing the reason for his decision in $t=10$, the subject is found to select alternative A so as to attribute (clearly) his defection to the past behaviour of his fellow members.

If we examine subsession 2's data having in mind these different theoretical predictions, the classification provided in Table 6.5 (and the corresponding distribution of types shown in Table 6.4) changes with regard to few subjects. First, all six always–cooperators of phase 3 who are categorised as strategic players in phase 6 of Table 6.5, can now be regarded as persuasive players; all these subjects expect their partners to defect in the last period, and all of them (as pointed out earlier) motivate their defective choice by picking out al-

ternative *A*. Likewise, subject 1, who is classified as strategic player in phase 4 and in phase 6 of Table 6.5, becomes—if we allow experience to matter—a persuasive agent in both phases. Table 6.5's classification, finally, changes with reference to subject 8 and subject 13: both of them can now be categorised as persuasive players (rather than as strategic players) in phase 6. For all the other subjects, the classification remains unaltered. Thus, data analysis carried out by regarding subsession 2 as a continuation of subsession 1 confirms the importance of learning about the others' behaviour (which leads people to converge towards the fully–rational Nash equilibrium) as well as the strong effect of the 'restart' (which induces people to be more cooperative after a break stopping the continuity of their decisions).

6.7 Concluding discussion

In this chapter, I wished to assess the relative success of various models of cooperative behaviour in explaining this experimental data. My main aim was to verify if and to what extent my hypothesis of persuasive behaviour can explain previously inexplicable cooperation.

The data led us to some important findings. First, the results provide clear evidence of heterogeneity with respect to motivations for cooperating not only across individuals, but also, within the same individual, across phases. Most of the subjects are found to modify their player type according to the supergame played. The change in the behavioural rule followed by a same individual subject over the six phases can be attributed to the different *mrs* which characterised phases as well as to the experience gained by the subject as the experiment proceeded.

The great importance of previous experience to people's behaviour is a second major finding of this experiment. A crossed study of the same individual's behaviour in each phase of the two identical subsessions included in my experiment clearly reveals that most subjects modify, in the second subsession, their behaviour and their attitude towards the others' interests, in the light of the experience gained in the first subsession. In many cases, we observe that as subjects gain knowledge of the other players' type, their behavioural pattern converges towards the non–cooperative, fully–rational Nash equilibrium. This indicates that learning is at work in this experiment, although not learning in the sense of Andreoni (1988) (i.e., learning about the incentives of the game and the dominant–strategy equilibrium), but only learning about the types of the playing partners.

As for the ability of the rival hypotheses of cooperative behaviour to explain the data, in order to determine how well each of them performs in this experiment, let us consider the percentage of subjects that, over all six phases, follow the classification criteria identified for each hypothesis so as to fall, in Tables 6.2 and 6.4, into the row corresponding to that hypothesis. We found that, out of a total of 144 individuals (24 subjects × 6 phases), less than 1 percent are of the reciprocal type; 5.5 percent are classified as Kantian/altruist players; 21 percent behave as predicted by the strategic hypothesis; 3.5 percent are persuaded agent; and, finally, 14 percent act in accordance with my hypothesis of persuasive behaviour. If we consider phases 4, 5 and 6 as continuations of phases 1, 2 and 3 respectively—so as to alter Table 6.4 with respect to some entries—the percentage of persuasive players increases to 20%, while that of strategic players decreases to 14%.

Thus, my data show that a quite reasonable percentage of the observed cooperative behaviour which cannot be explained by any of the previous theories is explicable with my new hypothesis. The study reported here is therefore encouraging for a theory of persuasion, but still preliminary. Further analysis must be undertaken before more decisive conclusions can be reached. Next chapter is aimed to provide such a deeper analysis.

Chapter 7

A deeper experimental analysis: towards more decisive conclusions

7.1 Introduction

Guttman (1986), Isaac and Walker (1992), Andreoni (1993), Chan et al. (1994), and Palfrey and Prisbey (1996, 1997), among other authors, have criticised the design where the equilibrium lies on the boundary of the strategy set. They claim that subjects in the experiments might make mistakes when they take their decisions and, in the corner–solution situation, mistakes necessarily lead to non–zero contributions and therefore to excess giving. This, according to them, yields to an overstatement of the importance of altruism. As an alternative, these authors propose reward structures that produce interior equilibria. This approach allows mistakes by subjects to manifest themselves as either excessive giving or excessive non–giving. However, it has been argued,[1] these designs introduce scope for potentially confounding effects.

Since the main purpose of the study carried out in this chapter is to test the relevance of persuasion in an environment more complicated than that used in the previous chapter, two different public good situations are here analysed. In one, both the Nash non–dominant–strategy equilibrium and the social optimum lie in the interior of the strategy space; I shall refer to this as the *basic game*. In the other (the so called *separation game*), the social optimum does not change, while the Nash equilibrium is put in the corner of the strategy space and becomes a dominant–strategy solution.

[1] See, for instance, Andreoni (1993); Keser (1996); and Sefton and Steinberg (1996).

A similar design allows me to verify if an agent tries to induce his 'selfish' fellow members to play in accordance with collective rationality (as predicted by the persuasive behaviour hypothesis) when it is not easy to perceive all the benefits connected with this decision, and to investigate how the agent's behaviour changes when the social optimum is still an interior solution but the source of confusion related to the subtlety of the Nash equilibrium is eliminated.

7.2 The voluntary contribution mechanism environments

My discussion starts with a description of the voluntary contribution mechanism decision environments utilised in the experiment presented in this chapter.

7.2.1 The *basic* non linear game

The basic game used in this experiment belongs to the class of voluntary contribution mechanism games that induce a unique interior Nash equilibrium.

In it, subjects played in groups of three. Each subject was endowed with 9 tokens, and each had to decide how many of these tokens he wanted to contribute to a public good. The payoffs of representative player i $(\forall i = 1, 2, 3)$ were generated from the Stone–Geary utility function (5.8):

$$U(g_i, y) = (w - g_i - a)^{\alpha}(y + b)^{1-\alpha}.$$

As pointed out in Chapter 5 (Subsection (5.2.1)), for $-\frac{\alpha}{1-\alpha}(3w + b) < a < w - \frac{\alpha}{1-\alpha}b$, $b > 0$ and $0 < \alpha < 1$ utility function (5.8) generates interior solutions.

In the experiment, the following parameter values were chosen: $w = 13$, $a = -6.73$, $b = 7$ and $\alpha = 0.58$. Thus, as can be verified, the unique non–dominant Nash equilibrium is for each subject to invest 2 tokens, and the symmetric social optimum is where each subject invests 7 tokens. The two interesting outcomes are therefore put at a symmetric distance from the boundaries of the strategy set and are to a great extent separated the one from the other.

In addition, the Nash equilibrium is unique in individual as well as in group–total donations: only the symmetric selection $g_i = 2$ for all i, and not any combination of individual contributions that results in the aggregate contribution of 6, constitutes an equilibrium. Thus, in my basic game, if subjects are found contributing, this excessive giving cannot be due to a coordination problem. In fact, if the equilibrium was unique in total, but not individual, donations, there would be multiple equilibria and subjects may not focus on the symmet-

ric selection; each subject might choose a strategy consistent with his preferred equilibrium, resulting in a strategy combination that is not an equilibrium. My reward structure does indeed eliminate this coordination problem.

However, the lack of a dominant strategy makes the game more difficult to understand: subjects may be unable to calculate their Nash equilibrium strategy or they may find the concept too subtle. Subjects must still conjecture about the behaviour of the others in order to compute their optimal response. An incorrect calculation or a conjecture that others may not play their Nash equilibrium strategies would lead to departures from the intended equilibrium, and it is in this sense that the structure has been regarded as potentially confusing to subjects.

I would add, however, that settings with interior solutions make also the social optimum less transparent and compelling to subjects who either may be incapable to understand the mutual benefits associated with it or may feel uncertain about their partners' ability to understand.

7.2.2 The *separation* game

My separation game eliminates the source of confusion related to the Nash solution: it exhibits, indeed, a unique dominant–strategy equilibrium.

The parameters of the payoff structure were chosen so that the basic and the separation games shared some important features; under both games, subjects were given the same endowment of 9 tokens and the socially optimal contributions were identical. Nevertheless, in the separation game, equilibrium aggregate contributions, subjects' earnings from the equilibrium as well as from the social optimum, and subjects' monetary penalties for a deviation from the equilibrium were different.

Specifically, in the separation game, subject i's ($\forall i = 1, 2, 3$) *net* reward structure was the following:

$$U(g_i, y) = \begin{cases} (w - g_i - a)^{\alpha}(y + b)^{1-\alpha} + 16, & \text{if } g_i = 0; \quad (1) \\ \frac{(w-g_i-a)^{\alpha}(y+b)^{1-\alpha}}{2}, & \text{if } g_i > 0. \quad (2) \end{cases} \tag{7.1}$$

This means that, in the separation game, (for any combination of the others' contributions) the individual reward from contributing *nothing* was 16 units greater than the corresponding individual reward in the basic game; while the individual reward from contributing *something* was half of that in the basic game in the sense that, although the monetary benefits generated from the public good were the same in both games, in the separation game, a subject

who contributed a positive amount kept for himself only half of these benefits and gave the remaining half to a *charitable institution.*

These changes are not trivial. Modification (7.1.1) transforms the game with a simple equilibrium (of investing 2 tokens in the public good) to a game with a dominant–strategy equilibrium (of contributing zero), and virtually eliminates subjects' confusion over the subtleties of computing a Nash equilibrium. Modification (7.1.2) transforms the game where subjects derive all the benefits from the production of the public good to a game where such benefits must be shared with a charity, and—*if we assume that people support public goods out of self–interest* (i.e., because of the earnings that they themselves can receive from them)—such a modification should eliminate any incentive to contribute.

It may be argued that, in the separation game, people contribute the amount of equilibrium because their confusion decreases in comparison with that in the basic game, and not because their incentives to contribute are removed. If such an argument is true, in the basic game, we should observe, levels of contributions above as well as below the equilibrium given that, in it, subjects' understanding of the Nash solution is supposed to be less. On the other hand, if, in the basic game, contributions are biased *above* the equilibrium, whilst, in the separation game, they correspond to the equilibrium, this would mean that in the former game people have a reason for over–contributing which disappears in the latter game.

Let us suppose, however, that the modifications included in the separation game were the following:

$$U(g_i, y) = \begin{cases} (w - g_i - a)^{\alpha}(y + b)^{1-\alpha}, & \text{if } g_i = 0, 1, 2; \\ \frac{(w-g_i-a)^{\alpha}(y+b)^{1-\alpha}}{2}, & \text{if } g_i = 3, ..., 9. \end{cases}$$

In this case, the reasons for a persuasive player's over–contributions are eliminated, and the equilibrium still lies in the interior of the strategy space. Nevertheless, a similar reward structure seems to advise people that 2 is a 'crucial' contribution, and this may help them to sort the Nash solution out in the basic game. Bringing into participants' mind the game's solutions is not my intention: the basic game is, in fact, designed in order to test the persuasion hypothesis in a potentially confounding setting.

7.3 Experimental design: parameters and procedures

The computerised experiment was run at the Center for Experimental Economics (EXEC) at the University of York (UK).[2] It was organised in three sessions with 24 subjects for each session (thus, a total of 72 people took part in my experiment).

Each session was constructed around the non linear public good games described in Subsections (7.2.1) and (7.2.2), and consisted of three supergames.

After having played a first *basic supergame*, in which the basic game was repeated 10 times, participants were randomly re–assigned to a new group of three people[3] for playing a *separation supergame* which consisted of 10–repetitions of the separation game. Then, at the end of it, new groups of three played a further ten–period *basic* supergame.[4]

Subjects were informed that, at the start of each supergame, they would be randomly assigned to a group of 3 players and that the group–composition would remain constant during each supergame. Although they could recognise the other participants in the room, the subjects did not know the identity of the other individuals in their group.

By following Andreoni's (1993) study, rather than telling the participants the exact functional forms (5.8) and (7.1) (which they might not understand), I presented the games in tabular forms with player i's own contributions as rows, the sum of others' contributions as columns, and the various $U(g_i, y)$ as entries. The payoff matrixes shown to subjects are reported in Tables B.1 and B.2 in Appendix B.

In order to make it clear that (for any combination of the others' contributions) each token invested in the public good would return the same total earnings under both the games, I decided to present the payoffs in the separation game as shown in Table B.2; this matrix differs from Table B.1 only in the first row, which corresponds to zero contribution. The alternative would have been to show the subjects their own net payoffs and tell them that (given a positive contribution) the amount earned would be matched by a donation to charity. The presentation of Table B.2 was preferred because it points well out that the monetary benefits generated by the public good do not change and,

[2]Also for this experiment, the program software was developed by myself and by the EXEC programmer Norman Spivey.

[3]This groups' re–arrangement was aimed to eliminate any possibility of strategic play or motives like revenge and envy in the participants' behaviour.

[4]This repetition of the basic supergame with different groups was directed to investigate if, in this experiment (as in the previous one), the attitude of 'experienced' players towards the others is influenced by previous interactions.

therefore, that the socially optimal contributions are identical.

On the other hand, in order to avoid misinterpretation on the part of the subjects and mental arithmetic (division by 2) to derive net earnings, as soon as a subject took a contribution decision, the line in the matrix corresponding to that contribution appeared underlined on his screen, displaying the net payoffs that he could earn from that decision; the subject was then asked if he was happy with his decision and, if not, he had the possibility to re–take it.

Before the experiment started, subjects received a list of charities from which, at the end of the experiment, they were asked to pick out one, in case their output data revealed that (through positive contributions during the separation game) they helped any.

7.3.1 The endogenous questionnaire and the non questionnaire treatment

In this experiment, as in the previous one, participants were required to answer a questionnaire. Every period (after having taken a contribution decision and observed the contributions of their partners), they had to:
(1) provide the main reason for their own decision;
(2) guess the motivations behind both their fellow members' decision;
(3) predict how much the others would contribute in the upcoming period.[5]

They answered the first two questions by picking out—for each of them—one of *nine* different alternatives.[6] They had to type a number from 0 to 9 on their keyboard to state their expectations about the others' next decisions.

To provide them with incentives strong enough to answer the questionnaire seriously and honestly, players were compensated for right guesses of the other individuals' motivations and for accurate predictions. Indeed, each period, in addition to their earnings from the public good, subjects were rewarded with 50 francs[7] if they guessed correctly the motivations lying behind the last choice of both their partners. They received a further 50 francs if their predictions of the others' next decision turned out to be right.

The questionnaire stage and its payment scheme were included also in my first *simple* experiment. However, in a departure from this, here I run a control

[5]It is worth reminding that it is a peculiarity of my experiment eliciting subjects' beliefs in t for $(t+1)$, after they have observed the others' actual contributions in t. Also in this case, this procedure was followed in order to make subjects answer all three questions in a unique stage so as to divide each period into two parts only.

[6]The lists of alternatives provided to subjects are reported on pages 206 and 207 in Appendix B.

[7]The franc was the unit of experimental money. The exchange rate between franc and real money was 1 franc = 0.5 pence.

treatment in which the questionnaire was not administered.

In fact, the act of eliciting motivations, guesses about the others' motivations and expectations may influence subjects' behaviour by affecting their rate of cooperation. To check for this possibility, in one of the three sessions of this experiment, the questionnaire stage was removed, and each period subjects only made a contribution decision. I shall refer to this treatment as the *non–questionnaire treatment.* A comparison of the subjects' contributions in the settings with and without the questionnaire will reveal if the mere act of asking for questions influences their behaviour.

7.3.2 The communication treatment

Although throughout all my experiment communication by words was forbidden and players were not allowed to speak to each other, I introduced a condition in which participants could communicate the motivations for their own decisions; i.e., their answer to the first question of the submitted questionnaire.

Among the structural mechanisms capable of affecting decision–making in dilemma situations, the effect of communication has been well established. In reviewing several studies, Dawes (1980) points out that the effects of communication on 'group oriented' decision making are ubiquitous, i.e. present in all the examined works. Being confronted with an opinion from others appears to be more effective in promoting cooperative behaviour for subjects with a cooperative social motivation, than for those more individualistically or competitively motivated.

Hence, in this study, when the questionnaire stage is included, two different experimental scenarios are analysed. In one—the *communication treatment*—participants communicated (non–optionally) the reason why they contributed that specific number of tokens to each other. Under this treatment, each subject knew that, at the end of each round, his own motivation response would be transmitted to the screen of his playing partners and that the motivation of each of the others would appear on his own screen. The other scenario—the *non–communication treatment*—did not include the transmission of these motives.

It may be argued that the introduction of this kind of communication modifies the nature of the game and the equilibria set. However, since no binding contracts or formal agreements can be closed, this communication is simply *cheap–talk*: it does not affect at all the payoffs and the unique equilibrium.

Nevertheless, by comparing the choices in these different conditions, I can isolate the impact of the others' opinions on people's behaviour and I can verify if the type of communication I have used increases cooperation.

7.3.3 Subject pool

The subjects were drawn from a list made (and updated every year) by EXEC, which consists primarily of undergraduate and graduate students of the University of York who ask to be informed in advance about all the experiments going on. All the participants were volunteers recruited by mail–shot invitations.

Upon their arrival, subjects were each seated at a computer terminal. Next to the screen, they found the list of charities that they could help through their decisions as well as a copy of the instructions.[8] These instructions were also read aloud. The subjects were then given an opportunity to ask questions before individually going through additional, computerised instructions.

After each of them finished reading the instructions, participants were randomly arranged in eight groups of three, and played the first 10–period game (i.e., the first basic supergame).

During the experiment, at the end of each round, subjects were shown information (by means of a 'results table' displayed on their screen) on their own experimental earnings for the round just finished as well as on the choices and corresponding earnings of each of their partners. In the communication–treatment, along with this information, the motivation response of each of the two partners was also available on the subjects' screen. At the end of the separation supergame, each subject was informed about the total donation that he gave to charity.

It was explained that the decisions for each round were binding and that end–of–experiment rewards would be based on the sum of earnings from all rounds. Subjects were told that the unit of experimental money was the franc, and that the value of each franc was 0.5 pence. The average payoff, earned in about one hour, was approximately 13 English pounds in the communication treatment and 10 English pounds in the non–communication treatment. The non–questionnaire treatment lasted about half an hour and, in it, the average payoff was approximately 8 English pounds.

After having received their own payment and before leaving the laboratory, the subjects who donate positive amounts to charity were asked to select one charitable institution from the provided list.

[8]Complete copy of the instructions is reported in Appendix B.

7.4 Persuasion and other theories of cooperative behaviour

In this section I shall present what alternative theories of cooperative behaviour predict in the games used in my experiment.

The way in which the basic game should be played under various theories has been discussed in Chapter 5, where I used utility function (5.8)—which generates each subject's payoffs in the basic game—as an illustrative device. The predictions/restrictions derived in that context will now be recalled and exploited in order to construct the rival behavioural categories in the basic game. For each of these theories, its predictions/restrictions in the separation game will also be carefully examined.

Although it would be desirable to have the same set of theories open to investigation here and in Chapter 6, it is not possible to take into account the strategies hypothesis as an explanatory category of this experiment's data. First of all, restriction (5.35) (on the amount of tokens that a strategic player must be observed to contribute in the first two periods of my basic game) depends on the probability, δ, that the player attaches to the likelihood that his playing partners are irrational, which is an unknown variable. Furthermore, given the non–binary dimension of the basic game's strategy set, it is not clear how strategic behaviour evolves in relation to the others' contributions in the time interval from period 3 to period $T-1$. As stressed in Chapter 5, in a game where the strategy set is not binary, the theory does not specify the decision rule that a strategic player should follow when he intends to modify his contribution from one period to the next. Since in this study I discriminate among theories by exploiting the fact that they differ on whether and how an individual's own contributions are related to the others' behaviour, due to the impossibility of deriving the form of this relationship for a strategic subject in the basic game, this chapter will not include 'strategies' among the theories under investigation.

7.4.1 Persuasive behaviour theory

In the basic game, a persuasive strategy is concisely and completely defined by restriction (5.12) and by the two derivatives (5.16) and (5.17).

For each level of contribution $\bar{g}_p \in \{0, 1, 2, ..., 9\}$, restriction (5.12) allows the calculation of the maximum number of periods, μ, for which a persuader can constantly contribute $\bar{g}_p$ unreciprocated (i.e., without being followed by each of his partners). Substituting the parameters w, α, a and b, and the equilibrium amount g^* into (5.12), the relationship between $\bar{g}_p$ and μ turns

out to be as displayed in Table 7.1. This table shows the persuasive player's reciprocity–test–period, μ, corresponding to each admissible contribution level, $\bar{g}_p$.

Table 7.1: Value of the reciprocity–test–period μ corresponding to each possible contribution level $\bar{g}_p$ in the basic game.

$\bar{g}_p$	9	8	7	6	5	4	3	2	1	0
μ	0	0	1	2	3	5	6	7	9	10

Table 7.1 clearly indicates that the relationship between μ and $\bar{g}_p$ is inverse, which confirms the result obtained in Chapter 5 (Subsection (5.2.1)). Furthermore, given that we can identify a persuader only when his μ is at least 3,[9] Table 7.1 suggests that if we observe a subject to contribute an amount greater than 5 tokens, unreciprocated, for some consecutive periods, then the subject cannot be pursuing a persuasive strategy. This table shows, in fact, that only if $\bar{g}_p \leq 5$, then $\mu \geq 3$. Thus, in the basic game, if a subject is observed to contribute, unreciprocated: *a*) more than 5 tokens, *b*) 5 tokens for more than 3 periods, *c*) 4 tokens for more than 5 periods, and *d*) 3 tokens for more than 6 periods, then we can infer that the subject is not a persuader.

The two derivatives (5.16) and (5.17) represent two additional testable restrictions on persuasive behaviour: by regressing the contributions of a subject on the total contributions of his partners in the previous period, if the slope of the regression line does not change (from zero to positive) in period μ, this would mean that the subject is not following a persuasive behaviour.

As far as the separation game is concerned, in it, the benefits of the public good must be shared with a charitable organisation. This means that the persuasive agent's main incentive not to be selfish disappears. Indeed, it is neither an altruistic concern nor a moral obligation towards the others that induce a persuasive player to contribute more than the equilibrium. Instead, this decision is taken with the expectation of future better gains for himself. Such an expectation can never be fulfilled in the separation game, where the private benefits related with the public good are less than those connected with the equilibrium. In the separation game, no value of $\mu > 2$ exists to justify positive contributions on the part of a persuasive player who is assumed not to care for the charity. Therefore, in the separation game, the persuasive behaviour is observational equivalent to the wholly 'selfish' Nash behaviour.

[9] See Subsection (5.2.2) for clarification about this point.

7.4.2 Commitment theories

In the basic game, commitment theories predict that an individual will contribute throughout the game a constant amount equal to the social optimum (see Eq. (5.23)). As indicated in (5.24), this implies that, by regressing a Kantian player's contributions on the total contributions of his partners, the regression line must exhibit a slope coefficient equal to zero. Thus, if a subject is observed not to contribute always 7 (so that a regression between his own and his partners' contributions generates a slope coefficient different from zero), then we can infer that the subject is not acting in accordance with commitment theories.

Let us now consider the separation game. In it, in order to decide how much to contribute, a Kantian type solves the following problem:

$$\max_{g_{i,t}>0} \quad (w - g_{i,t} - a)^{\alpha}(y_t + b)^{1-\alpha}$$

subject to:

$$g_{k,t} = g_{i,t} \quad \forall k \neq i \quad \forall t \in [1, 10],$$

whose solution (as showed in Chapter 5) is $g^*_{c,t} = \hat{g} = 7$.

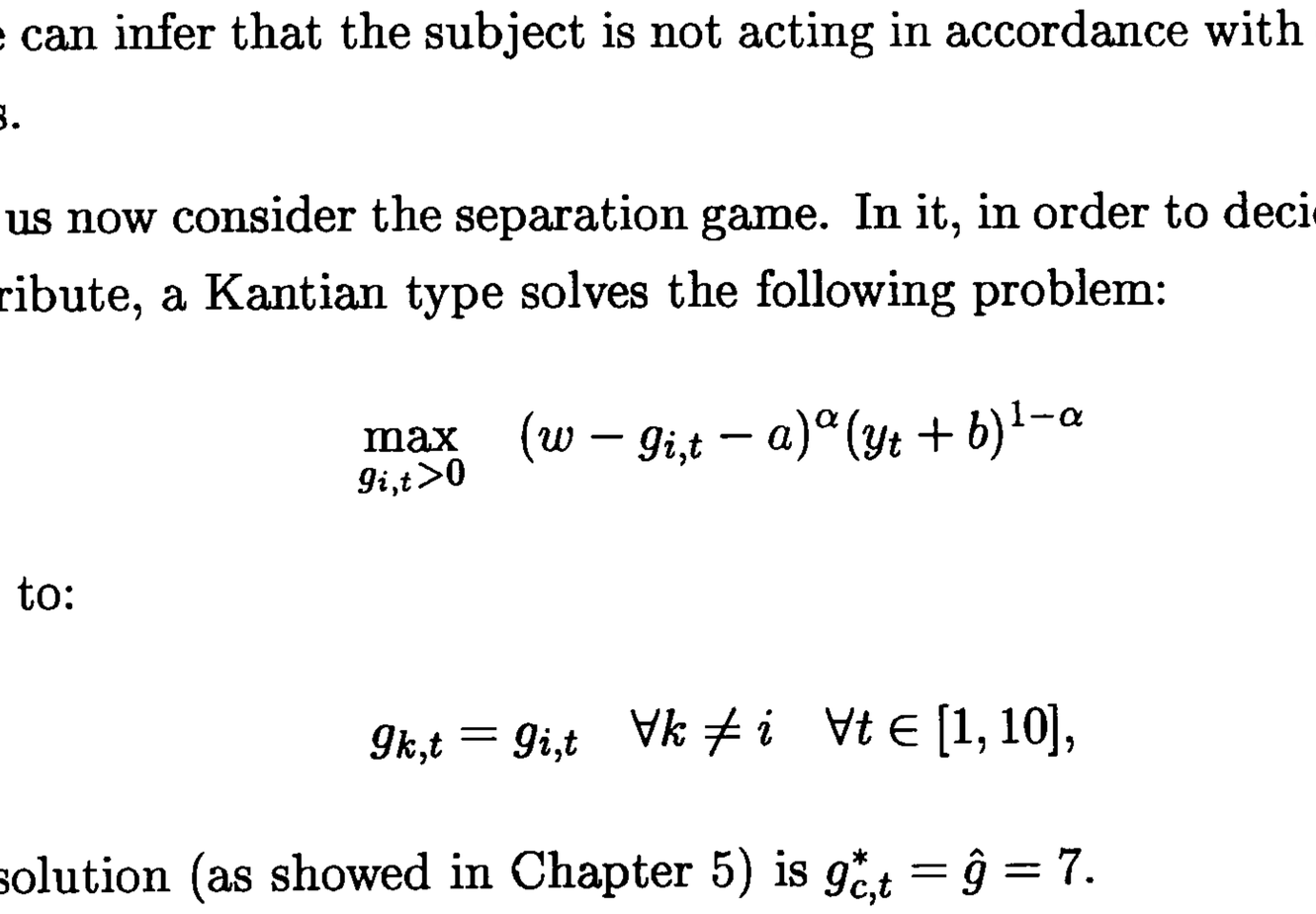


Figure 7.1: Kantian constrained utility in the separation game.

The latter, in the separation game, is no longer a global maximum, since the discontinuity of the utility function in $g_{i,t} = 0$ implies a further local maximum at zero. Nevertheless, by comparing the two local maxima (i.e., by comparing

the agent's utility in correspondence of each of them), we find that the absolute maximum still occurs at $g_{i,t} = 7$. This is illustrated by points M_1 and M_2 in Fig. 7.1, which graphs the Kantian constrained utility for all possible values of $g_{i,t}$. It is easy to see that both M_1 and M_2 represent extreme values, but i's utility is highest at M_2. Thus, also in the separation game, a Kantian player is predicted to contribute always 7, regardless of the others' behaviour.

7.4.3 Reciprocity theories

In Chapter 5, we saw that the moral obligation upon which a reciprocator bases his choice leads him to contribute more than the equilibrium only if everyone else in his group does so as well. In Subsection (5.5.2), I demonstrated that, in a setting with the Stone–Geary utility function used in the basic game, a reciprocal individual's contributions lie between the fully–rational Nash equilibrium of contributing 2 tokens and the symmetric social optimum of contributing 7 tokens, depending on the other group members' behaviour in the previous period. Derivative (5.25) states, in fact, that the relationship between a reciprocator's contribution in t and the sum of the others' contribution in $t-1$ is positive, which means that a reciprocator increases (decreases) his contribution in round t if, in $t-1$, it was below (above) the average of the others. Relationship (5.25) represents a useful testable restriction on reciprocal behaviour: by regressing a subject's own contributions on those of his partners, if the slope coefficient does not come up to be positive, then we can infer that the player is not a reciprocator.

In the separation game, a subject who acts in accordance with Sugden's reciprocity principle maximises utility function (7.1) subject to the constraint: $g_{i,t} \equiv min(g_c^*, g_{k,t-1})$, where g_c^* is the optimal level of contribution under commitment theories. If $min(g_c^*, g_{k,t-1}) = g_c^*$, agent i is obliged to contribute at least g_c^*, which implies $g_{r,t}^* = 7$. If $min(g_c^*, g_{k,t-1}) = g_{k,t-1} = 0$, agent i is obliged to contribute at least zero, i.e. the dominant–strategy Nash solution. On the other hand, it can never be $g_{r,t}^* > 7$, because the individual would be contributing more than he is obliged to. Thus, in the separation game, the optimal contribution of a reciprocal person varies from the dominant–strategy equilibrium of contributing 0 to the social optimum of contributing 7.

Furthermore, restriction (5.25) holds also in the separation game. Sugden proved, indeed, that the relationship given by (5.25) has a general validity: a reciprocator's contributions must be positively related to the contributions of his partners whatever the function used to model the subject's preferences.[10]

[10] Cf., Sugden (1984, p. 780).

7.4.4 Altruism theories

In the basic game, the way in which altruistic behaviour evolves in relation to the others' contributions is given by Eq. (5.30), which displays the altruist's reaction function. By differentiating such a function by the sum of the others' contributions we obtain a negative derivative, as shown by (5.31). This means that, in the basic game, we can classify a subject as an altruist only if a regression between his own and his partners' contributions presents a negative slope coefficient.

Let us now consider the separation game. The discontinuity of payoffs present in such a game originates a discontinuity in the altruist's reaction function whenever γ (the altruist's weight on the others' welfare) is such that $0.4 \leq \gamma \leq 0.7$. For values of γ lying in the latter interval, if we calculate the altruistic optimal contribution g_a^* as given by (5.30), we obtain that $V(g_a^*, y) > V(0, y)$ insofar as $\sum_{k \neq a} g_k$ does not exceed a critical value, $\xi(\gamma)$. For values of the sum of the others' contributions greater than $\xi(\gamma)$, the altruist results to be better off by contributing zero (the dominant–strategy 'selfish' Nash equilibrium) than by contributing g_a^*. This means that for specific values of γ and of $\sum_{k \neq a} g_k$, the local maximum at g_a^* is not longer a global maximum.

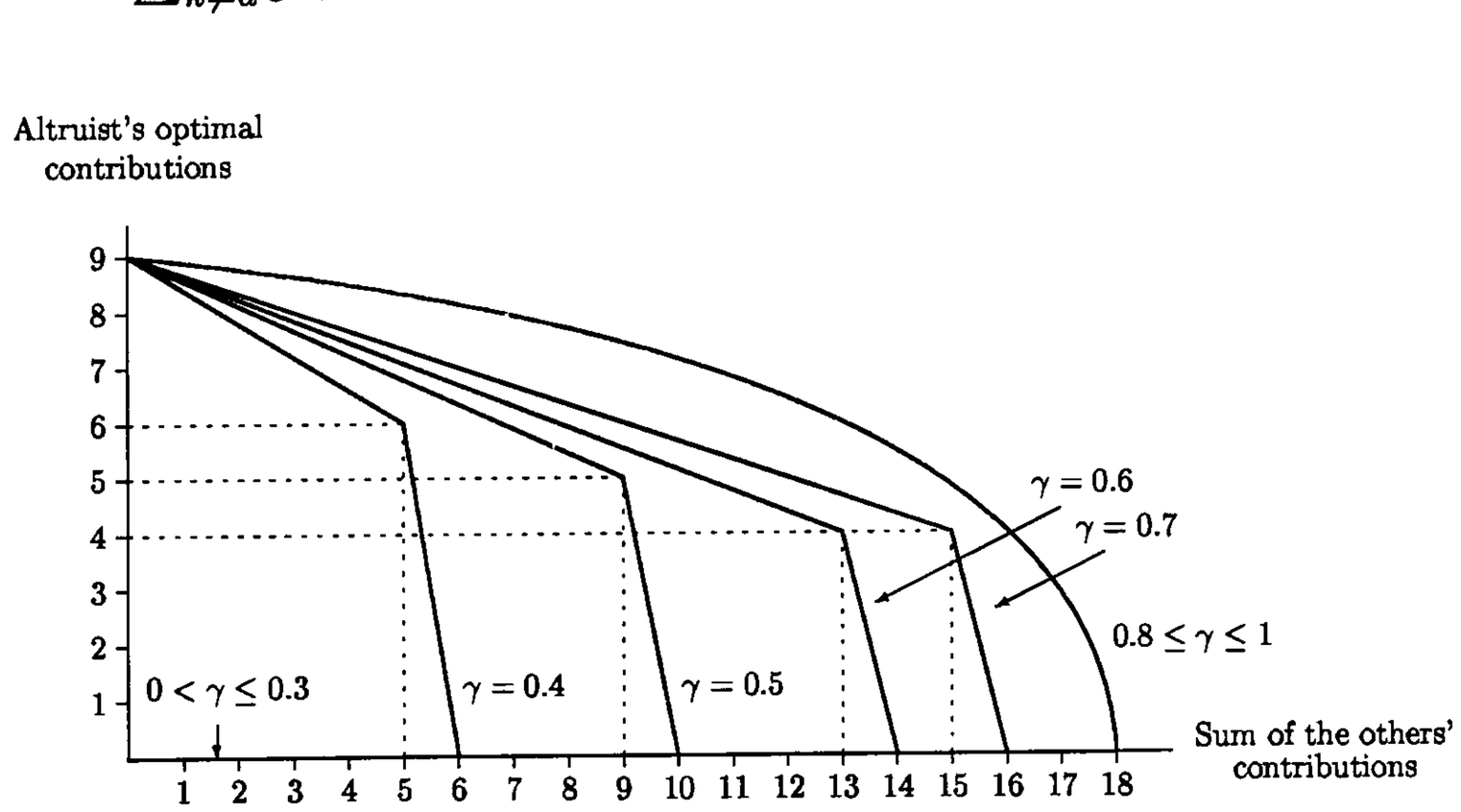


Figure 7.2: Altruist's reaction function in the separation game.

The behaviour of an altruist in the separation game depends, therefore, on his own γ. Fig. 7.2 graphs the altruist's reaction function for different values of γ. Specifically, we have the following.

a) If $0 < \gamma \leq 0.3$, the altruist is predicted not to contribute anything, whatever his partners' behaviour throughout the game; thus, for these values of γ, the global maximum is at zero.

b) If $0.4 \leq \gamma \leq 0.7$, the altruist's reaction function exhibits a discontinuity. Consider the following general statement: "If $\gamma = \gamma^*$, the altruist will contribute g_a^* if $\sum_{k \neq a} g_k < \xi(\gamma^*)$; otherwise, he will contribute 0". Then, if $\gamma^* = 0.4 \Rightarrow \xi(0.4) = 6$. If $\gamma^* = 0.5 \Rightarrow \xi(0.5) = 10$. If $\gamma^* = 0.6 \Rightarrow \xi(0.6) = 14$. If $\gamma^* = 0.7 \Rightarrow \xi(0.7) = 16$.

c) Finally, if $0.8 \leq \gamma \leq 1$, the altruist reaction function is that given by (5.30). In this case, the global maximum still occurs at g_a^*.

7.5 Aggregate results

My data are 2160 observed contributions to a public good (30 decisions for each of 72 subjects) collected from three separate sessions. Each session employed one of three treatments: *communication treatment* (henceforward CT), where the questionnaire was administered and subjects were able to transmit information about the motives of their contributions; *non–communication treatment* (henceforward NCT), in which the questionnaire was conducted, but the transmission of the subjects' motivations was not included; and *non–questionnaire treatment* (henceforward NQT), i.e., the treatment without any questions. Each session–treatment consisted of three supergames: the first and the third supergames were based on the basic game, while the second supergame was constructed around the separation game. Let *phase* be another word for supergame. Thus, throughout the following discussion, phases 1, 2 and 3 refer to the first basic supergame, the separation supergame and the second basic supergame, respectively.

Summary results are reported in figures 7.3, 7.4 and 7.5. Fig. 7.3 presents the individual average contributions separately for each phase and treatment. Fig. 7.4 displays the individual contributions associated with equilibrium, social optimum and their mean (4.5 tokens in phases 1 and 3, and 3.5 tokens in phase 2), and charts average contributions in the three treatments for each period of the three phases. Fig. 7.5 compares average contributions in the three phases for each treatment, and reports the individual contributions associated with the mean between equilibrium and social optimum in phases 1 and 3.

7.5.1 Treatment effects on aggregate behaviour

This subsection investigates if, in aggregate, there are treatment effects. Subsection 7.5.2 examines if average behaviour changes across phases.

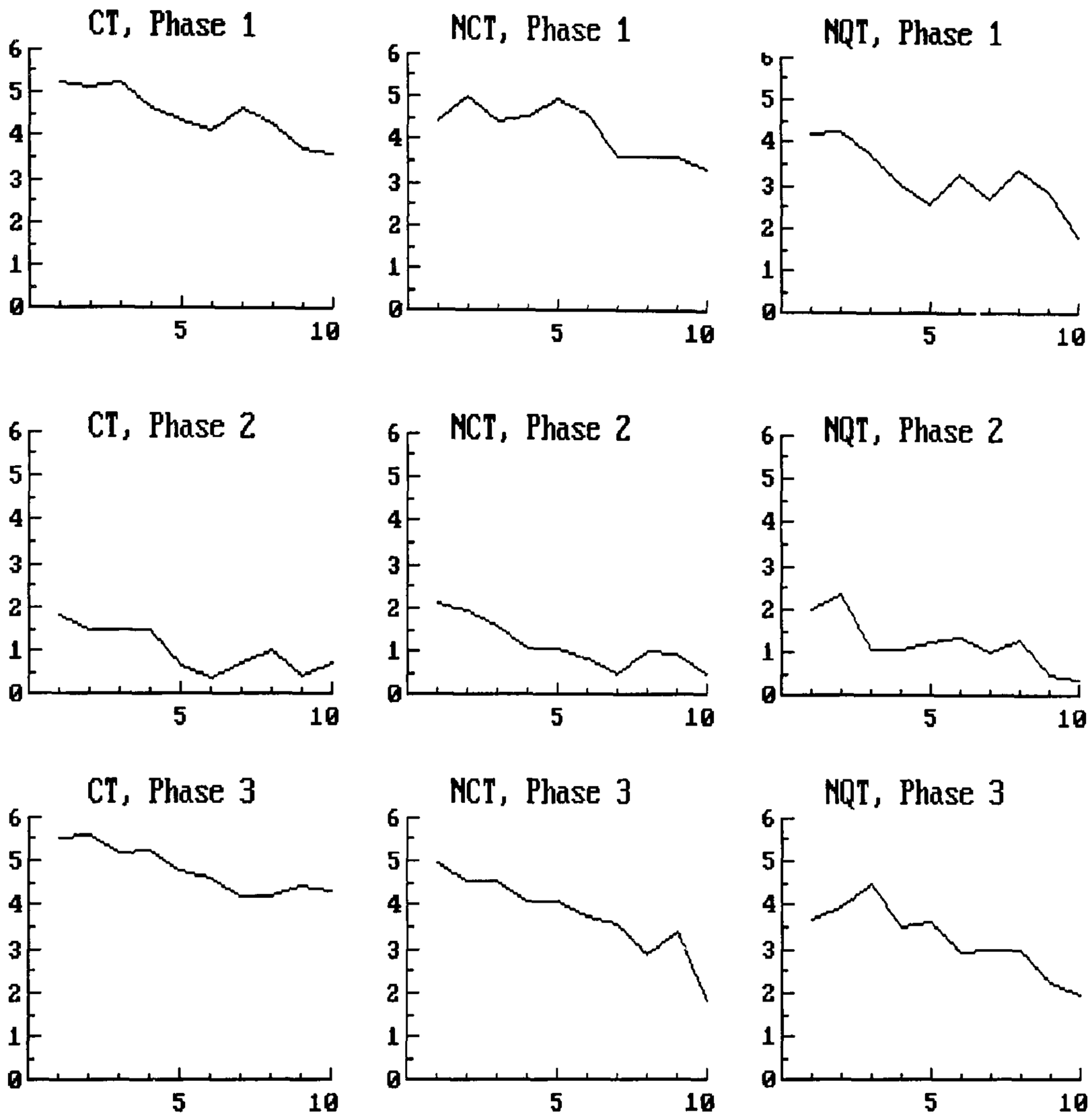


Figure 7.3: Average individual contributions displayed separately in the 3 phases of each treatment.

The effects of communication

The communication treatment was included to explore whether the possibility for the subjects to communicate the reasons why they contributed that specific number of tokens increases the incidence of cooperative choice. If it is true that being confronted with an opinion from others is more effective in promoting contributive behaviour for subjects with a cooperative social motivation, than for those more individualistically or competitively motivated, one should expect the rate of cooperation to be higher in the communication treatment than in the two treatments without communication.

Table 7.2 shows the differences in average contributions between the CT and each of the two treatments without communication over all ten rounds of

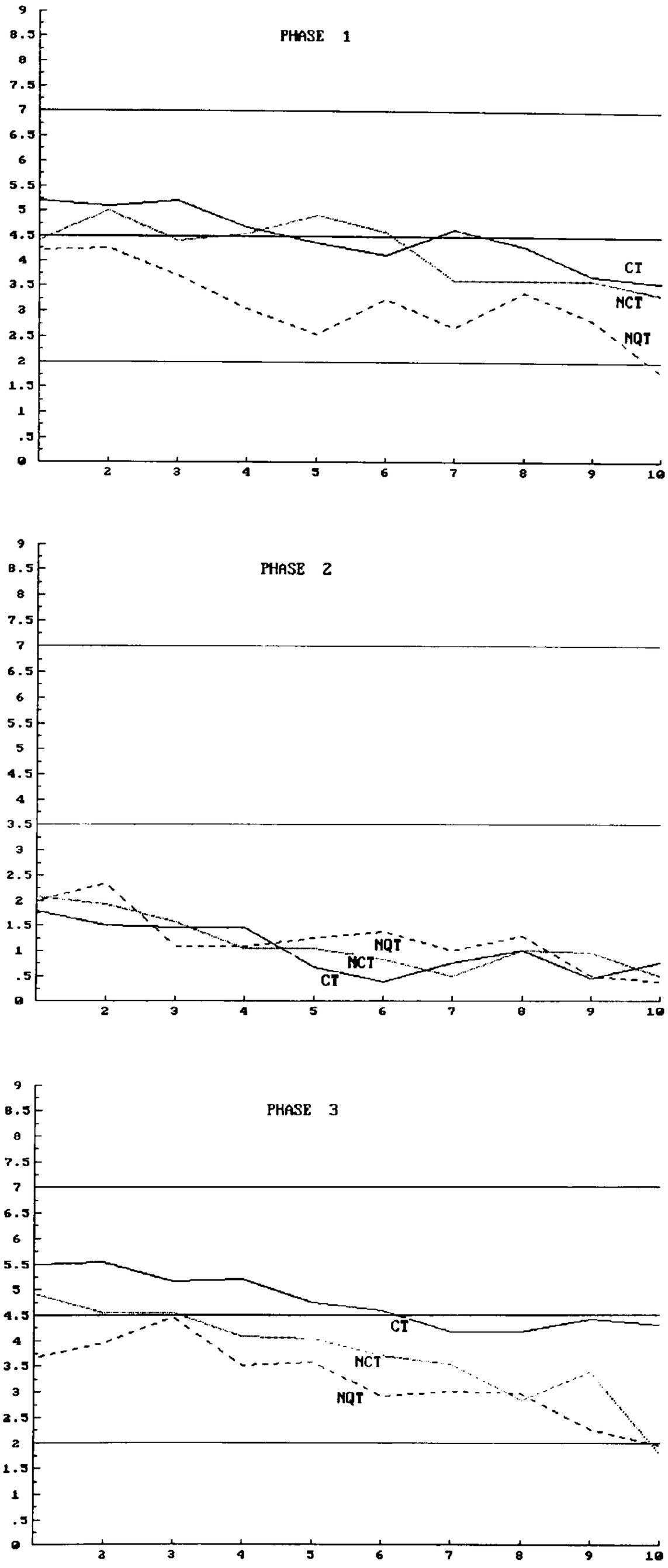


Figure 7.4: Average contributions in the three treatments for each phase.

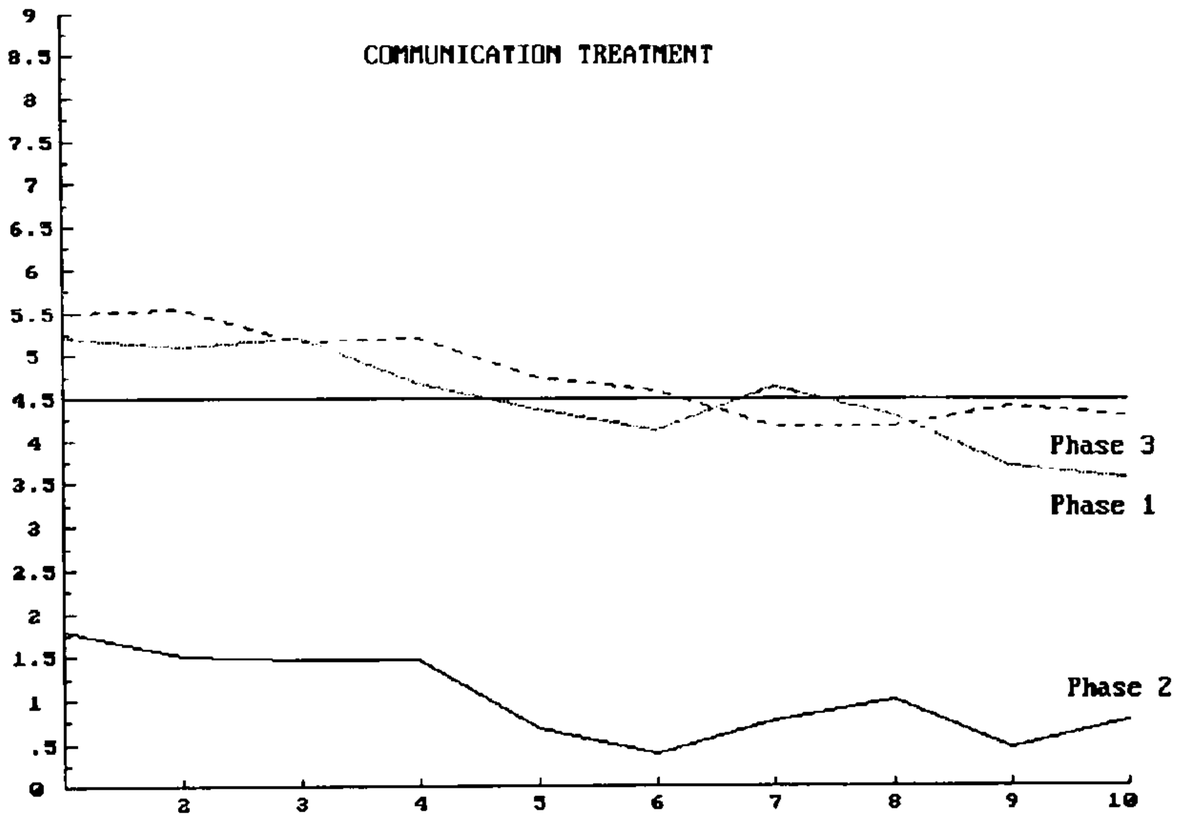


Figure 7.5: Average contributions in the three phases for each treatment.

each phase.

Table 7.2: Differences in individual average contributions between the CT and the two treatments without communication.

	PHASE 1		PHASE 2		PHASE 3	
	CT - NCT	CT - NQT	CT - NCT	CT - NQT	CT - NCT	CT - NQT
1	0.79	1.00	−0.29	−0.21	0.58	1.83
2	0.08	0.83	−0.42	−0.83	1	1.58
3	0.79	1.50	−0.13	0.38	0.63	0.71
4	0.13	1.63	0.42	0.38	1.13	1.71
5	−0.54	1.83	−0.38	−0.58	0.71	1.17
6	−0.46	0.87	−0.46	−1.00	0.88	1.67
7	1.00	1.96	0.25	−0.25	0.63	1.17
8	0.67	0.92	0.00	−0.29	1.33	1.21
9	0.08	0.88	−0.50	−0.04	1.04	2.17
10	0.25	1.79	0.25	0.38	2.50	2.33

First, let us compare the CT and the NCT. As can be seen from Fig. 7.4, in phases 1 and 3 average contributions under the CT are, most of the periods, above those under the NCT. Inspection of Table 7.2 reveals that this happens in 8 periods of phase 1 and over all phase 3. From examination of Fig. 7.5 appears that, while average contributions tend to fall with repetition under the NCT, they are almost steady around the mean between equilibrium and social optimum under the CT. Under the assumption of independent observations, if we pool the data over phases, and test—for each phase—the null hypothesis that the mean contributions of the two treatments are equal, we reject the null hypothesis for both phase 1 and phase 3 (at the 10% and at the 1% levels of significance, respectively).[11]

[11]Let μ_C and μ_{NC} be the respective means for the CT and the NCT. For each phase, I test the null hypothesis $H_0 : \mu_C - \mu_{NC} = 0$ against the alternative $H_1 : \mu_C - \mu_{NC} > 0$. The decision rule is to reject H_0 in favour of H_1 if

$$\frac{\overline{X}_C - \overline{X}_{NC}}{\sqrt{(s_C^2/n_C) + (s_{NC}^2/n_{NC})}} = \tilde{z}_{C,NC}(phase) > z_\alpha,$$

where $\overline{X}_C$ and $\overline{X}_{NC}$ are the observed sample means for the CT and the NCT respectively; s_C and s_{NC} are the respective sample standard deviations; $n_C = n_{NC} = 240$ is the number of observations; and $\tilde{z}_{C,NC}(phase)$ is the value obtained comparing the means of the CT and the NCT in a particular phase.

In phase 1, the following values are found: $\overline{X}_C = 4.49$, $\overline{X}_{NC} = 4.20$, $s_C = 2.55$ and $s_{NC} = 2.43$; so that $\tilde{z}_{C,NC}(1) = 1.28$. In phase 3, $\overline{X}_C = 4.78$, $\overline{X}_{NC} = 3.74$, $s_C = 2.60$ and $s_{NC} = 2.63$; therefore $\tilde{z}_{C,NC}(3) = 4.37$.

Thus, as far as the two phases constructed around the basic game are concerned, the kind of communication used in this experiment is found to affect people's behaviour in that it pushes them toward more cooperation.

Interestingly, this result changes drastically when we consider phase 2. Fig. 7.4 shows that, in such a phase, average contributions in the CT are most of the time below those in the NCT. Table 7.2 confirms that, in 6 out of the 10 periods, the proportions of tokens contributed to the public good is lower under the CT than under the NCT, and that, in one period, they are equal. The possibility of transmitting information does not seem to help cooperation when there are no 'private' benefits from it. Moreover, the test of no difference in individual mean contributions reveals that there is not a statistic significant difference between the CT and the NCT in phase 2.[12]

Similar results are obtained when we compare the CT with the NQT. People contribute significantly more to the public good under the CT than under the NQT in phases 1 and 3, i.e. when they keep all the benefits associated with this choice. Instead, when the benefits from investing own tokens in the public good must be shared with a charitable institution (as happens in phase 2), in 7 out of 10 periods, average contributions are less under the CT than under the NQT. In addition, the difference in mean contributions between treatments is found to be significant at the 1% level in phases 1 and 3, and insignificant in phase 2.[13]

All these findings lead to the following first observation.

Observation 7.1 *At the aggregate level, communication strongly affects cooperation in the two 10–period phases where subjects retain all the monetary benefits generated from the production of the public good. Instead, when these benefits must be shared with charity, average contributions under the CT are not significantly different from those under the two treatments without communication.*

The effects of the questionnaire

The treatment without any question was introduced to investigate whether the act of eliciting beliefs about the others' motivations and actions influences a subject's own contributions.

---

[12] In phase 2, $\overline{X}_C = 1.02$ and $s_C = 2.36$; $\overline{X}_{NC} = 1.15$, and $s_{NC} = 2.26$. Hence, $\tilde{z}_{C,NC}(2) = -0.59$.

[13] In particular, in phase 1, $\overline{X}_{NQ} = 3.17$ and $s_{NQ} = 2.21$; in phase 2, $\overline{X}_{NQ} = 1.23$ and $s_{NQ} = 2.42$; in phase 3, $\overline{X}_{NQ} = 3.23$ and $s_{NQ} = 2.16$. Therefore, $\tilde{z}_{C,NQ}(1) = 6.07$, $\tilde{z}_{C,NQ}(2) = -0.95$, and $\tilde{z}_{C,NQ}(3) = 7.12$.

From examination of Fig. 7.4 it appears that average contributions under the NQT are most of the time below those under the other two treatments, especially as far as phases 1 and 3 are concerned. Fig. 7.5 clearly reveals that, while average contributions under the NQT lie always, in all three phases, below the corresponding means, this does not apply to the average contributions of the CT and NCT in phases 1 and 3.

Table 7.3 reports the differences in average contributions between the NCT and the NQT over all ten rounds of each phase. The differences between the CT and the NQT are depicted in Table 7.2.

Table 7.3: Differences in individual average contributions between the NCT and the NQT.

	PHASE 1	**PHASE 2**	**PHASE 3**
	NCT - NQT	NCT - NQT	NCT - NQT
1	0.21	0.08	1.25
2	0.75	−0.42	0.58
3	0.71	0.50	0.08
4	1.50	−0.04	0.58
5	2.38	−0.21	0.46
6	1.33	−0.54	0.79
7	0.96	−0.50	0.54
8	0.25	−0.29	−0.13
9	0.79	0.46	1.13
10	1.54	0.13	−0.17

As stressed in the previous subsection, if phases 1 and 3 are taken into account, average contributions are always higher under the CT than under the NQT, and the mean contributions of the two treatments are significantly different. Instead, when phase 2 is considered, in 7 periods, average contributions are higher under the NQT than under the CT, and the difference in mean contributions is not significant.

Inspection of Table 7.3 reveals that the differences in average contributions between the NQT and the NCT follow similar patterns: in phases 1 and 3, subjects contribute (on average) less when their beliefs and motivations are not elicited than when these are elicited; in phase 2, subjects contribute more when beliefs and motivations are not elicited than when they are elicited. Moreover, the difference in mean contributions between the NCT and the NQT is found

to be significant at 1% level in phases 1 and 3, and not significant in phase 2.[14]

Thus, we can state the following.

Observation 7.2 *Asking subjects to motivate their own choices and to think about the motivations and the future decisions of their partners affects, at the aggregate level, their behaviour in the two phases constructed around the basic game in that it pushes them towards more cooperation.*

Interestingly, this finding contrasts with those shown by Croson (1998a, 1998b) and by the psychological literature.[15] Indeed, in these earlier experiments, contributions are lower when expectations are elicited.

7.5.2 Interior versus boundary equilibria

I turn, here, to a direct comparison of subjects' average behaviour in the three phases of each treatment. From inspection of Fig. 7.4 and Fig. 7.5 it seems that, at the aggregate level, subjects behave differently according to phase, especially under the NCT.

In phase 1, and under the two treatments where the questionnaire was administered, average contributions are steady around the mean of equilibrium and social optimum (i.e., 4.5 tokens) at least in early decision rounds.

In phase 2, in all treatments, averages are always below the corresponding mean (i.e., 3.5 tokens).

In phase 3, averages lie systematically above the mean in the first 5 periods and systematically below it in the last 5 periods under the CT. This downward trend takes place earlier and it is much more sharp under the NCT.

Table 7.4 reports the results of the test of no difference in individual mean contributions between phases, under each treatment.[16] As can be seen, average contributions are significantly different when we compare phase 1 and phase 3 with phase 2, under all treatments; and when we compare phase 1 with phase 3, under the NCT.

These findings lead to the following two observations.

Observation 7.3 *In phase 2, where the dominant strategy is no provision and where the benefits from the public good must be shared with a charity, contributions fall (on average) far short of the social optimum in every period.*

---

[14] Specifically, the following values are found: $\widetilde{z}_{NC,NQ}(1) = 4.87$, $\widetilde{z}_{NC,NQ}(2) = -0.39$, and $\widetilde{z}_{NC,NQ}(3) = 2.34$.

[15] Cf., e.g., Shafir and Tversky (1992); and Shafir (1994).

[16] In Table 7.4, $\widetilde{z}_{i,j}$ is the value of the statistics obtained if we compare phase i with phase j.

Table 7.4: z–statistics for the differences between phases under each treatment.

	CT	NCT	NQT
$\widetilde{z}_{1,2}$	15.5^{***}	14.2^{***}	9.16^{***}
$\widetilde{z}_{1,3}$	-1.23	2.00^{*}	-0.29
$\widetilde{z}_{3,2}$	16.6^{***}	11.6^{***}	9.59^{***}

$^{***}\ p < 0.001$ $^{*}\ p < 0.05$

Observation 7.4 *Under the treatment in which the questionnaire was administered but no communication was included, people are (on average) less cooperative in phase 3 than in phase 1.*

The strength of Obs. 7.3 and Obs. 7.4 can be refined if we take into consideration the percentages of Nash equilibrium strategies that are observed in each phase, by separating according to treatment. Table 7.5 displays these percentages.

Table 7.5: Percentages of Nash equilibrium strategies observed in each phase of each treatment.

	NCT	CT	NQT
Phase 1	17.5	5.00	24.2
Phase 2	74.2	71.2	73.3
Phase 3	21.7	8.75	29.6

Consistent with the data on average contributions, we observe a high number of subjects following the strong free–riding strategy in phase 2. The proportions of 'egoists' is, on the contrary, quite low in phases 1 and 3, although it increases from the first to the third phase.

Because the settings with an interior–non–dominant–equilibrium are likely to be less transparent and compelling to subjects, the apparent more cooperative behaviour observed in them might be due to people's confusion over this too subtle solution. If deviations from the Nash equilibrium are simply randomly distributed errors attributable to learning, then, in phases 1 and 3, we would observe contribution levels below as well as above the Nash predictions. On the other hand, if there remains an asymmetry toward over investment to

the public good, this would suggest that there is a more complex explanation for behaviour biased above the non–cooperative equilibrium, towards the social optimum.

To give more insight into behaviour in the interior–Nash settings, let us consider the percentages of subjects who contribute less than the amount associated with the equilibrium—by following Sefton and Steinberg's study (1996) I shall refer to these people as *too greedy for their own good*—and the percentages of subjects contributing more than the amount of equilibrium. People contributing above the equilibrium can be further classified in *a*) *easy riders*, those that contribute 3–6 tokens; *b*) *welfare maximisers*, those that give 7 tokens, and *c*) *martyrs*, those that contribute 8 or 9 tokens. Table 7.6 shows these percentages separately for each treatment.

Table 7.6: Percentages of non–equilibrium strategies.

PHASE 1

	Too greedy for their own good	*Easy–riders*	*Welfare maximisers*	*Martyrs*
NCT	15.4	35.4	30.0	1.70
CT	22.5	35.4	37.1	0.00
NQT	22.9	40.8	9.20	2.90

PHASE 3

NCT	24.2	24.5	29.2	0.40
CT	23.3	21.2	46.7	0.00
NQT	20.3	37.6	12.1	0.40

Examination of Table 7.6 leads to the following observation.

Observation 7.5 *Across all treatments, there is a pronounced bias above the equilibrium in both phase 1 and phase 3.*

If we take into account the percentages of subjects whose contributions lie in a set of 'relatively' narrow bands around the equilibrium (i.e., we compare the proportions of easy–riders with those of subjects too greedy for their own good) Obs. 7.5 is more evident. In phase 1, there are systematically more easy–riders than too greedy people under all treatments. In phase 3, the relationship is reversed in the CT (where, however, a high percentage of subjects is found contributing 7 tokens), whilst it is kept under the NCT and the NQT.

How might we explain all these figures and observations?

The behavioural differences that, at the aggregate level, subjects exhibit in phase 2 in comparison with phases 1 and 3 suggest that the theories which predict no change in behaviour across phases do not play a main part in explaining this experimental data. At odds with the predictions of these theories, people are found to be significantly less cooperative when they cannot keep all 'private' benefits associated with the public good but must share them with a charitable organisation.

On the other hand, the way in which contributions are dispersed around the Nash predictions in the two basic supergames tells us that deviations from the equilibrium are not simply transitional errors due to subjects' confusion over the subtleties of computing an interior solution. Consistent with earlier patterns of allocations to group account, I find in fact that, in the interior–Nash supergames, decisions are systematically biased above the equilibrium.

These results suggest that there must be a more complex explanation for people's contributing behaviour.

An analysis of behaviour across individuals—which will be carried out in the next subsection—should help to throw more light on the reasons why people voluntarily contribute to public goods.

7.6 Individual results

I now turn to the analysis of behaviour across individuals and I discriminate among competing theories by exploiting the fact that they make different predictions about the relationship between a subject's own contributions and the contributions of his partners. Thus, for each of the 72 participants in the experiment, I estimate the following OLS regression:

$$g_{i,t} = c + s \sum_{k \neq i} g_{k,t-1} \tag{7.2}$$

where the only explanatory variable for subject i's contributions in t are the total contributions of i's fellow members in $t-1$.

Since (7.2) is a dynamic equation, I cannot take into account the first round's data. Consequently, the sample size is 9 observations per phase. However, as far as the two phases constructed around the basic game are concerned (i.e., phases 1 and 3), it may be possible to increase the number of observations available per person, if the difference in the coefficients on others' contributions between these two phases was insignificant. Thus, before proceeding to the estimation (for each single phase) of the coefficients' values that best fit an individual's

data, I test for the equality of these coefficients in phase 1 and phase 3.

7.6.1 The phase effect at individual level

For each subject, I estimate a model which uses the observations of both phase 1 and phase 3, and which includes a dummy variable D taking value 0 in the first phase and 1 in the third phase. That is, for each participant, I calculate the following regression equation:

$$g_{i,t} = c + (C - c)D + s\sum_{k \neq i} g_{k,t} + (S - s)D\sum_{k \neq i} g_{k,t} \tag{7.3}$$

where:

$$D = \begin{cases} 0 & \text{for } t = 1, ..., 10 \text{ (phase 1)} \\ 1 & \text{for } t = 21, ..., 30 \text{ (phase 3).} \end{cases}$$

The coefficients $(C - c)$ and $(S - s)$ give the difference, respectively, in the intercepts and in the coefficients on others' contributions between phases. If, in a regression, the value of the coefficients turns out to be insignificant, this allows me to pool the data over the two phases and have, regarding that regression, 18 observations.

Notice that this test looks only at the way in which a subject's behaviour changes with respect to the others' contributions. Therefore, even if the subject's level of contributions are different across phases, the two phases will be regarded as equal and data will be pooled provided that the relationship between own contributions and others' contributions is the same.

The results of this analysis are reported in Table 7.7, that shows, treatment by treatment, the number of individuals for which the value of the coefficient was found significant.

Table 7.7: Number of cases in which the phase effect is significant.

NCT	CT	NQT
1	2	4

As can be seen, the phase effect is statistically significant for only one subject under the NCT, for two subjects under the CT, and for four subjects under the NQT. This means that, for all the other participants, it is possible to pool the data over the two phases.

7.6.2 The cutoff period θ

As pointed out in Subsection (7.4), commitment and reciprocity theories predict that the sign of the slope s in Eq. (7.2) should be the same across all three phases (despite the discontinuity of payoffs present in the separation game). According to commitment theories, the slope should be zero in each phase; according to reciprocity theories, it should be positive in each phase.

In contrast with these two theories, the other two theories investigated in this study prescribe that subjects behave differently depending on the phase they are involved in, so that the sign of s in Eq. (7.2) varies according to phase.

In the case of altruism, the slope of the regression line is predicted to be negative in the two phases constructed around the basic game. Instead, in the phase based on the separation game, the slope is negative—for all combinations of the others' contributions—only if γ (the weight that the altruist puts on the others' welfare) is greater than 0.7; if γ lies between 0.4 and 0.7, the slope is negative insofar as the others' contributions do not exceed a specific value (otherwise the altruist contributes zero); if γ assumes values less than 0.4, the global maximum of the altruist's reaction function is at zero, which means that—whatever the others' contributions—the altruist does not contribute anything throughout phase 2.

As far as a persuasive player is concerned, he is expected to contribute always zero in phase 2; whilst, in phases 1 and 3, he is predicted to contribute a constant positive amount till period μ and to reciprocate the others' behaviour thereafter. Thus, under the persuasive behaviour hypothesis, the relationship is zero in phase 2, while it turns from zero to positive in phase 1 and in phase 3. This implies that, if my hypothesis applies, by splitting phase 1 and phase 3 in two parts and by estimating separate models for each part, the sign of the relationship in the two models should be different.

Thus, in order to test the predictions of these competing theories, in addition to regression (7.2), I estimate also the following model:

$$g_{i,t} = \underline{c}_\theta + (\bar{c}_\theta - \underline{c}_\theta)D_\theta + \underline{s}_\theta \sum_{k \neq i} g_{k,t-1} + (\bar{s}_\theta - \underline{s}_\theta) \sum_{k \neq i} g_{k,t-1} D_\theta \qquad (7.4)$$

where D_θ is a 'round' dummy variable taking value 0 for the first θ rounds and 1 for the last $(10 - \theta)$ rounds; $\theta = 2, ..., 9$ if phases 1 and 3 data can be pooled or $\theta = 3, ..., 8$ if the regressions must be estimated for each phase separately;[17]

[17] Since in order to draw a regression line at least two points are necessary, the dummy D_θ must take value zero (or one) for at least two rounds. This implies that, when the regression is calculated for a single phase, θ has to vary from 3 to 8 given that the first round's contribution decisions are not taken into account. On the other hand, when data can be pooled over phases

$\underline{c}_\theta$ and $\underline{s}_\theta$ are, respectively, the intercept term and the slope coefficient for a regression calculated in the first θ periods of a phase; and $\bar{c}_\theta$ and $\bar{s}_\theta$ denote the intercept term and the slope coefficient for a regression calculated in the remaining $(10 - \theta)$ periods.

Since θ is a variable taking value either 2 to 9 or 3 to 8, model (7.4) yields 8 or 6 different regression equations, each of which divides a phase into two intervals with θ as cutoff point.

7.6.3 Comparing decisions with others' behaviour

In this subsection, for each of the 72 participants in the experiment, I estimate the coefficients of the 8 or 6 regressions generated by model (7.4), and (as the split of a phase in two parts may not exist) I estimate also the coefficients of regression (7.2), either over all periods of each phase (if the phase effect is significant), or with data pooled over all periods of phases 1 and 3 (if the phase effect is insignificant). Among all these regressions, I select, for each subject, the one that minimises Akaike's information criterion.

This kind of analysis allows me to verify, for each individual, *if* and *when* the slope coefficient changes its sign during a particular phase. I can hence identify which of the competing theories under examination (if any) is the best predictor of the subject's behaviour.

According to the characteristics of the selected model, I allocate individuals into different groups. Specifically, a subject will be classified as:

- a *reciprocator*, if, in any phase, the regression that best fits his data is Eq. (7.2) and, in it, $s > 0$; or, even if there is a split, $\underline{s}_\theta$ and $\bar{s}_\theta$ are both positive;

- a *Kantian* individual, if, in any phase, Eq. (7.2) is selected and, in it, $s = 0$;

- an *altruist*, if, in any phase, Eq. (7.2) turns out to be the best regression and, in it, $s < 0$; or, even if there is a split, $\underline{s}_\theta$ and $\bar{s}_\theta$ are both negative. On the other hand, an agent who is found to contribute always nothing to the public good in phase 2 might also be an altruist; precisely, he would be an altruist who puts a weight less than 0.4 on the others' welfare;

- a *persuasive* player, if, in phases 1 and 3, the selected regression has a split

1 and 3, if $\theta = 2$, D_2 equals zero in the second round of both phases, and if $\theta = 9$, D_9 equals one in the last round of both phases; hence, in these cases, there exist two points through which the regression line can pass.

in $\theta \geq 3$ and presents $\underline{c}_\theta \leq 5$,[18] $\underline{s}_\theta = 0$ and $\bar{s}_\theta > 0$. On the other hand, a persuasive player must contribute nothing to the public good throughout phase 2.

Notice that I allow both a reciprocator and an altruist to have a 'split' period within a same phase (i.e., in both cases, I allow for $\underline{s}_\theta$ and $\bar{s}_\theta$ to differ); this means that a subject, while preserving the form of the relationship, becomes more/less reactive to the others' behaviour. This kind of change is permitted by both altruism and reciprocity theories, which require that the *sign* of the slope (not its absolute value) must remain the same throughout a supergame.

Whenever a subject's selected regression does not exhibit any of the characteristics mentioned above, the subject will not be arranged in any behavioural category but will be considered *unclassifiable.*

It is worth emphasising that, in two cases, the regression coefficients cannot be estimated: 1) when the individual exhibits constant own contributions, which amounts to having no variation in the dependent variable; and 2) when the sum of the others' contributions is constant, which means that it is impossible to detect the effects of it on the dependent variable. As for the first issue, (in a parallel to the individual data analysis carried out in the previous chapter) individuals who contribute a constant amount greater than the equilibrium throughout a phase will be considered *always-contributors*; on the other hand, those who contribute a constant amount equal to the equilibrium will be classified as players of the *Nash type.*

The relationship under the NCT

Phases 1 and 3 Under the NCT, the phase effect is not significant for 23 out of a total of 24 participants (see Table 7.7). The regression coefficients and the best θ (if any) of each of these 23 subjects are shown in Table 7.8.[19]

Inspection of Table 7.8 reveals the following.

a) Four subjects (coherent with altruism theories) exhibit negative slope coefficients over all periods.

b) Three more subjects (consistent with reciprocity theories) present positive slope coefficients over all periods.

[18] A similar restriction derives from (5.12) and from the corresponding values displayed in Table 7.1. These values suggest that if a player is observed to contribute an amount greater than 5 tokens (unreciprocated) for more than 2 periods, then he cannot be of the persuasive type.

[19] In this table as well as in all the following ones, the decimal numbers with more than two zero digits are rounded to zero.

Table 7.8: Characteristics of each subject's pooled regression under the NCT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
5	Altruist	5	1.05	3.59	−0.10	−0.14
15	Altruist	5	12.5**	2.66*	−0.81	−0.03
7	Altruist		$c = 5.99$		$s = -0.10$	
23	Altruist		$c = 3.21^{***}$		$s = -0.17^{**}$	
1	Reciprocal	4	1.02	−3.91	0.14	1.11*
2	Reciprocal		$c = 1.30$		$s = 0.30$	
17	Reciprocal	5	−0.43	0.74	0.84*	0.30
3	Persuasive	3	4.00*	0.61	0.00	0.44
4	Unclassified	7	12.9**	1.28**	−1.01*	0.47*
6	Unclassified	2	13.0***	1.03***	−0.80*	0.34**
8	Unclassified	3	7.53	0.28	−0.10	0.39
9	Unclassified	9	3.63***	4.67	0.10	−0.33*
10	Unclassified	6	7.00***	0.14***	0.00	0.64**
12	Unclassified	7	7.00*	−9.33***	0.00	0.19*
13	Unclassified	5	7.00***	−1.13***	0.00	0.52**
14	Unclassified	8	1.66	7.50*	0.17	−1.00*
16	Unclassified	5	7.00***	0.65***	0.00	0.43***
18	Unclassified	6	6.02**	−1.96**	0.00	0.44*
19	Unclassified	5	7.00***	3.20***	0.00	0.27***
20	Unclassified	4	7.00***	0.72***	0.00	0.39*
22	Unclassified	4	3.37***	2.00	−0.22*	0.00
24	Unclassified	2	−32.0***	0.77***	3.00***	0.00***
21	Always–Contrib.		$g_{21,t} = 7 \quad \forall t$			

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$

c) Only one subject is found to behave in accordance with the persuasive behaviour hypothesis: he contributes 4 tokens until the third period and then he reciprocates the others' contributions.

d) One individual contributes 7 tokens in all periods and, as a consequence, he is considered an always–contributor.

e) Finally, a large number of participants are unclassifiable. Among these unclassifiable patterns of behaviour, we observe some homogeneity which is

worth emphasising. For instance, a few subjects are found to contribute 7 tokens (i.e., the socially optimal amount), *unreciprocated*, for some consecutive periods before their slope coefficient becomes positive.[20]

The only person who exhibits a significant phase dummy is of the Nash type in phase 1; while he counts as unclassifiable in phase 3 (in this phase, he starts contributing 2 tokens only after the fourth round). Table 7.9 reports the characteristics of his regression equation.

Table 7.9: Characteristics of the regression for the subject whose phase effect is significant under the NCT.

PHASE 1

Subject	*Type*	
11	Nash	$g_{11,t} = 2 \qquad \forall t$

PHASE 3

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
11	Unclassified	4	5.65^{***}	2.00^{***}	-0.41^{***}	0.00^{***}

*** $p < 0.001$

Phase 2 In the second phase of the NCT, for 3 subjects, the explanatory variable is constantly equal to zero, which implies that we cannot detect its effects on the dependent variable. The characteristics of the selected regression of each of the remaining 21 participants are reported in Table 7.10.

Examination of this table reveals the following.

a) One subject is altruist. He fell in this category also in phases 1 and 3; thus, we can infer that his own γ is greater than 0.7.

b) Consistent with reciprocity arguments, a positive relationship over all phase is observed in 4 cases.

c) Three participants do not present readily interpretable data and, consequently, are considered unclassifiable.

d) Finally, an insufficient variation in the dependent variable is detected for 13 subjects, who are found contributing always zero. Accordingly, they are all classified as Nash types. Among them, we find the only persuasive player of phases 1 and 3 as well as one of the four altruists of phases 1 and 3. The subject for which the latter is observed might be categorised as an altruist whose γ is less than 0.4, rather than as a type Nash player.

[20] Examples of this kind of behaviour are offered by subjects no. 10, 12, 13, 16, 19 and 20.

Table 7.10: Characteristics of each subject's selected regression in the second phase of the NCT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
5	Altruist		$c = 1.99$		$s = -0.22$	
4	Reciprocal	4	-1.49	-4.00	0.27	2.00
6	Reciprocal		$c = -0.43$		$s = 0.64^{***}$	
8	Reciprocal		$c = 2.18$		$s = 0.29$	
14	Reciprocal		$c = 0.37$		$s = 0.52^{*}$	
1	Unclassified	7	0.00	4.50	0.00	-0.50
7	Unclassified	6	7.00	0.38	0.00	0.11
13	Unclassified	8	5.00^{***}	2.00^{**}	-0.10	0.50^{*}
3	Nash		$g_{3,t} = 0 \quad \forall t$			
9	Nash		$g_{9,t} = 0 \quad \forall t$			
10	Nash		$g_{10,t} = 0 \quad \forall t$			
11	Nash		$g_{11,t} = 0 \quad \forall t$			
12	Nash		$g_{12,t} = 0 \quad \forall t$			
16	Nash		$g_{15,t} = 0 \quad \forall t$			
18	Nash		$g_{18,t} = 0 \quad \forall t$			
19	Nash		$g_{19,t} = 0$ in $t = 2, ..., 10$			
20	Nash		$g_{20,t} = 0 \quad \forall t$			
21	Nash		$g_{21,t} = 0$ in $t = 2, ..., 10$			
22	Nash		$g_{22,t} = 0 \quad \forall t$			
23	Nash		$g_{23,t} = 0 \quad \forall t$			
24	Nash		$g_{24,t} = 0 \quad \forall t$			

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$

The relationship under the CT

Phases 1 and 3 Under the CT, we can estimate 22 pooled regressions. In one of them, the explanatory variable equals 14 in all periods so that we cannot discern its effects on the dependent variable. The remaining 21 subjects exhibit regression coefficients as reported in Table 7.11.

Inspection of this table shows the following.

a) In only one case the slope is always negative, as predicted by altruism theories.

b) In one more case the slope is always positive, as predicted by reciprocity

Table 7.11: Characteristics of each subject's pooled regression under the CT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
7	Altruist	3	-5.11	2.23	-0.89	-0.05
18	Reciprocal	2	-13.7^{**}	-6.52	1.33^{***}	0.94
4	Persuasive	3	5.00	5.39	0.00	0.40
10	Persuasive	3	5.00	-3.67^{*}	0.00	1.22^{*}
1	Unclassified	4	7.00^{*}	0.36^{*}	0.00	0.44
2	Unclassified	3	8.68^{***}	1.00^{***}	-0.73^{***}	0.00^{***}
3	Unclassified	4	-8.34	4.38^{*}	1.34^{*}	-0.29^{*}
5	Unclassified	7	4.36^{***}	5.73	0.00	-0.82^{*}
8	Unclassified	2	6.00^{***}	4.81	-0.17	0.10
9	Unclassified	3	-17.0	3.13^{*}	1.50^{*}	-0.26^{*}
11	Unclassified	4	7.88^{***}	0.75^{***}	-0.63^{***}	0.10^{***}
12	Unclassified	6	7.00^{***}	-2.76^{**}	0.00	1.09^{**}
13	Unclassified	7	7.00^{***}	-0.23^{***}	0.00	0.42^{**}
16	Unclassified	9	-0.38	2.33	0.51^{***}	-0.17^{*}
14	Always–Contributor		$g_{14,t} = 5 \;\; \forall t = 1, ..., 10;$ $g_{14,t} = 7 \;\; \forall t = 21, ..., 30$			
15	Always–Contributor		$g_{15,t} = 7 \quad \forall t = 1, ..., 10, 21, ..., 30$			
17	Always–Contributor		$g_{17,t} = 5 \;\; \forall t = 1, ..., 10;$ $g_{17,t} = 7 \;\; \forall t = 21, ..., 30$			
19	Always–Contributor		$g_{19,t} = 7 \quad \forall t = 1, ..., 10, 21, ..., 30$			
22	Always–Contributor		$g_{22,t} = 5 \;\; \forall t = 1, ..., 10;$ $g_{22,t} = 7 \;\; \forall t = 22, ..., 30$			
23	Always–Contributor		$g_{23,t} = 7 \quad \forall t = 1, ..., 10, 21, ..., 30$			
24	Always–Contributor		$g_{24,t} = 7 \quad \forall t = 1, ..., 10, 21, ..., 30$			

$^{***}\, p < 0.001$ $^{**}\, p < 0.01$ $^{*}\, p < 0.05$

theories.

c) Two subjects can be considered persuasive players. Before their slope coefficient becomes positive, both invest 5 tokens, unreciprocated, for three periods.

d) An insufficient variation in the dependent variable is found in seven cases, suggesting that these subjects should be classified as always–contributors. In

particular, out of these seven subjects, four invest 7 throughout both phases; three invest 5 throughout phase 1 and 7 throughout phase 3.

e) Finally, ten persons remain unclassified. A careful look at some of these unclassifiable behavioural patterns reveals that also under the CT (as under the NCT) there are a few subjects who contribute 7 tokens, unreciprocated, for some consecutive periods before starting reciprocating the others' contributions.[21]

As for the two subjects who show a significant phase dummy, the features of their regressions are described in Table 7.12.

Table 7.12: Characteristics of each subject's regression when the phase effect is significant under the CT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
PHASE 1						
6	Unclassified	6	0.00	7.52*	0.00	−0.37
21	Unclassified	3	1.00	1.29	0.00	−0.10
PHASE 3						
6	Unclassified	3	7.00**	2.29*	0.00	0.03
21	Reciprocal		$c = -2.12$		$s = 0.37$*	

** $p < 0.01$ * $p < 0.05$

Phase 2 By turning to the analysis of the relationship in the second phase of the CT, the regression of two participants present the regressor equals zero over all phase. Table 7.13 reports details of the regression of each of the remaining 22 participants.

The following results are obtained.

a) One person behaves as an altruist.

b) Three subjects present an always positive slope coefficient, in accordance with reciprocity theories.

c) Three participants are unclassifiable.

d) Finally, 15 subjects are found to be of the Nash type since they contribute always zero. The two persuasive persons as well as the altruist of phases 1 and 3 fall in this category. The latter might, therefore, be classified as an altruist with $\gamma < 0.4$.

[21] The subjects who exhibit this behavioural pattern are subjects no. 1, 12, and 13.

Table 7.13: Characteristics of each subject's selected regression in the second phase of the CT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
18	Altruist		$c = 2.61^*$		$s = -0.21$	
8	Reciprocal		$c = 1.00$		$s = 0.71^*$	
19	Reciprocal		$c = -5.18$		$s = 1.31$	
16	Reciprocal	6	-0.38	2.83	0.34	0.81
6	Unclassified	3	-17.5^{***}	0.00	3.50	0.00^{***}
15	Unclassified	4	7.00^*	5.40	0.00	0.23
20	Unclassified	4	7.00^*	0.72^*	0.00	0.02
1	Nash		$g_{1,t} = 0 \quad \forall t$			
2	Nash		$g_{2,t} = 0 \quad \forall t$			
3	Nash		$g_{3,t} = 0 \quad \forall t$			
4	Nash		$g_{4,t} = 0 \quad \forall t$			
5	Nash		$g_{5,t} = 0 \quad \forall t$			
7	Nash		$g_{7,t} = 0 \quad \forall t$			
9	Nash		$g_{9,t} = 0 \quad \forall t$			
10	Nash		$g_{10,t} = 0 \quad \forall t$			
11	Nash		$g_{11,t} = 0 \quad \forall t$			
12	Nash		$g_{12,t} = 0$ in $t = 2, ..., 10$			
13	Nash		$g_{13,t} = 0 \quad \forall t$			
21	Nash		$g_{21,t} = 0 \quad \forall t$			
22	Nash		$g_{22,t} = 0 \quad \forall t$			
23	Nash		$g_{23,t} = 0 \quad \forall t$			
24	Nash		$g_{24,t} = 0 \quad \forall t$			

*** $p < 0.001$ * $p < 0.05$

The relationship under the NQT

Phases 1 and 3 Under the NQT, it is possible to pool the data over phases 1 and 3 for 20 subjects, whose regression coefficients and best θ (if any) are shown in Table 7.14.

Inspection of this table reveals the following.

a) Four participants present negative coefficients over all periods, in accordance with altruism theories.

Table 7.14: Characteristics of each subject's pooled regression under the NQT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
4	Altruist	8	5.27***	2.29	−0.29**	−0.41
10	Altruist	7	9.14***	9.86	−0.60**	−1.36
24	Altruist	2	14.0*	2.25*	−1.00	−0.10
12	Altruist		$c = 3.51^*$		$s = -0.08$	
15	Reciprocal	2	−9.33	2.15	1.33	0.02
11	Reciprocal	5	2.48***	3.69*	0.17**	0.11
7	Reciprocal	6	4.85***	1.23*	0.04	0.36
3	Reciprocal	9	2.10*	0.00	0.14	0.20
1	Reciprocal		$c = -2.46$		$s = 0.46$	
17	Persuasive	3	5.00*	3.76	0.00	0.46
21	Nash		$g_{21,t} = 2 \quad \forall t$			
8	Unclassified	3	10.0**	2.15*	−1.25*	0.02*
23	Unclassified	3	15.5	−0.08	−0.82	0.45
19	Unclassified	8	6.96**	−7.67*	−0.47	2.33*
20	Unclassified	5	3.19*	7.05	0.15	−0.43*
5	Unclassified	7	−0.15	4.93	0.31	−0.46
14	Unclassified	9	2.35***	5.33	0.03	−0.67
9	Unclassified	4	7.00*	1.33	0.00	0.26
2	Unclassified	5	7.00***	−6.90***	0.00	2.00**
13	Unclassified	7	3.67***	2.00	−0.23	0.00

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$

b) Five subjects exhibit positive coefficients over all periods, in accordance with reciprocity theories.

c) One person is persuasive: he contributes 5 tokens, unreciprocated, for 3 periods and then he reciprocates the others' behaviour.

d) One subject is found contributing always the amount of tokens associated with the equilibrium. He can be therefore considered of the Nash type.

e) Finally, the regression of nine participants present features such that these participants cannot be arranged in any category. Two out of these nine unclassifiable persons are found to contribute the socially optimal amount, unreciprocated, until their cutoff period θ and to reciprocate the others' contri-

butions thereafter.

As for the 4 participants whose phase dummy is found to be significant, the regression coefficients and the best θ of each of them are reported in Table 7.15.

Table 7.15: Characteristics of each subject's regression when the phase effect is significant under the NQT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
PHASE 1						
6	Unclassified	7	10.5^{**}	-44.0^{**}	-1.15^{**}	5.50^{**}
16	Unclassified	5	7.00^{*}	2.00	-1.00	0.00
18	Unclassified	4	7.00	0.62	0.00	0.38
22	Altruist		$c = 2.30^{***}$		$s = -0.10^{*}$	
PHASE 3						
6	Reciprocal	7	3.80	-5.14	-0.07	-0.43^{*}
16	Unclassified	5	2.05^{**}	2.00	0.29^{***}	0.00
18	Unclassified	3	7.00	2.00^{***}	0.00	0.00
22	Unclassified	5	7.00^{*}	-2.00	0.00	0.73

$^{***}\ p < 0.001$ $^{**}\ p < 0.01$ $^{*}\ p < 0.05$

Phase 2 In the second phase of the NQT, the regression of two participants exhibit a zero regressor throughout the phase. The coefficients and the best θ of the remaining 22 subjects are depicted in Table 7.16.

The following can be observed.

a) For six participants the slope of the regression line is always positive, as predicted by theories of reciprocity.

b) Six more participants are unclassifiable.

c) The remaining ten subjects contribute always zero. They are, therefore, categorised as Nash types. Among these always–defectors we find the persuader as well as one of the altruists of phases 1 and 3.

Discussion

A few observations stand out from this study.

First of all, it shows not only that there is a considerable variety of behaviour across individuals, but also that there are a substantial number of

Table 7.16: Characteristics of each subject's selected regression in the second phase of the NQT.

Subject	*Type*	*Best* θ	$\underline{c}_\theta$	$\bar{c}_\theta$	$\underline{s}_\theta$	$\bar{s}_\theta$
10	Reciprocal	3	-12.0^*	0.11^*	4.00^{**}	0.26^{**}
23	Reciprocal	4	-17.5	2.00^*	4.50^*	0.50
15	Reciprocal	8	6.20^{**}	-3.00^*	0.13	4.00^*
7	Reciprocal		$c = 0.98$		$s = 0.44^*$	
11	Reciprocal		$c = 0.17$		$s = 0.23$	
19	Reciprocal		$c = -3.47$		$s = 0.89$	
1	Unclassified	4	4.17^{**}	-0.21^*	-2.50^*	0.70^{**}
6	Unclassified	5	5.00^*	1.25	-0.55	0.53^{**}
22	Unclassified	8	1.16	-0.60	-0.07	0.60
8	Unclassified	7	2.27	8.85	1.21	-1.60^*
5	Unclassified	5	-0.89	0.00	1.07^{**}	0.00^{**}
13	Unclassified	8	-0.34	0.00	0.74^{***}	0.00
9	Nash		$g_{9,t} = 0$ in $t = 2, ..., 10$			
20	Nash		$g_{20,t} = 0$ in $t = 2, ..., 10$			
2	Nash		$g_{2,t} = 0 \quad \forall t$			
3	Nash		$g_{3,t} = 0 \quad \forall t$			
4	Nash		$g_{4,t} = 0 \quad \forall t$			
14	Nash		$g_{14,t} = 0 \quad \forall t$			
16	Nash		$g_{16,t} = 0 \quad \forall t$			
17	Nash		$g_{17,t} = 0 \quad \forall t$			
18	Nash		$g_{18,t} = 0 \quad \forall t$			
21	Nash		$g_{21,t} = 0 \quad \forall t$			

$^{***}\ p < 0.001$ $^{**}\ p < 0.01$ $^{*}\ p < 0.05$

subjects who modify their behaviour and their attitude towards the others during a supergame. This is especially true for phases 1 and 3, where most of the regressions that best fit an individual's data exhibit a cutoff period.

In these two phases, under the NCT, only 1 out of the 21 subjects whose data can be pooled shows an earlier zero–later positive relationship as predicted by the persuasive behaviour hypothesis; 3 subjects exhibit a positive relationship as predicted by reciprocity theories; and 4 subjects present a negative relationship

as predicted by altruism theories. No one shows a zero coefficient over all periods as predicted by commitment theories. When the phase effect is significant, we find one more reciprocator in phase 1, and one more altruist in phase 3.

Under the CT, a conspicuous number of subjects is represented by always–contributors; such a result is consistent with the data on average contributions. As for the ability of the competing theories to explain the CT's data, none of them appear to fare well in this treatment. Indeed, when data can be pooled, 2 subjects out of 21 behave in accordance with the persuasive behaviour hypothesis, 1 in accordance with reciprocity theories, and another one in accordance with altruism theories. Again, no one acts according to commitment theories. When regressions are estimated separately for each phase, we find one more reciprocator in phase 3.

As far as the NQT is concerned, an interesting feature of the data is that, under this treatment, in phases 1 and 3, we do not find any always–contributor but, rather, we observe one player of the Nash type. This contrasts with the results obtained under the other two treatments and confirms that, in my experiment, individuals are less inclined to contribute when they are simply required to state a contribution decision and are not 'forced' to think about their own and the others' behaviour. As pointed out on p. 166, this finding differs from those reported by Croson (1998b) and in the psychological literature; in these earlier experiments, in fact, the act of eliciting beliefs significantly decreases contribution levels and cooperation.

By turning to the phase based on the separation game, in it, under all treatments, no one behaves as a Kantian individual, and very few subjects behave as reciprocators or altruists. Instead, a remarkable proportion of participants is represented by always–defectors. Some of these always–defectors are found to be altruists in the two phases constructed around the basic game. This suggests that, in the classifications of the subjects which refer to phase 2, these participants might be categorised as altruists with $\gamma < 0.4$ rather than as Nash types.

In addition, in phase 2, we observe more subjects of the Nash type in the two treatments with the questionnaire than in the treatment without questions. Under the NCT and the CT, 13 and 15 persons (respectively) contribute the equilibrium amount throughout the phase while, under the NQT, only 10 players behave as Nash types. My second phase's data appear, hence, to be in line with Croson's result. It is necessary, however, to take into account that, in contrast with the previously cited experimental studies, in phase 2 of my experiment subjects do not get all the benefits connected with the public good

but they must share them with a charitable organisation. Such a difference between the second phase of my experiment and the earlier experimental works makes a direct comparison between them difficult.

A last aspect of the data which deserves to be underlined concerns the significance of the estimated coefficients: a substantial number of them, indeed, is found to be statistically insignificant. This may be due to the few number of observations available per person as well as to the amount of noise in the data. However, even if many coefficients lack statistical significance, from an economic point of view (which matters here) they are significant.

7.6.4 Comparing an individual's expectations with the others' contributions

A further study that the data from this experiment allows us to do concerns the relationship between each subject's expectations and the actual patterns of contributions of his fellow members.

In order to verify if subjects form expectations on the basis of what they observe, for each of the 48 participants in the two treatments in which beliefs were elicited, I calculate the following OLS regression:

$$E_{i,t}\left[\sum_{k\neq i} g_{k,t+1}\right] = \alpha_0 + \alpha_1 \sum_{k\neq i} g_{k,t}, \tag{7.5}$$

where the only explanatory variable for subject i's expectations in t are the observed total contributions of his partners.

If, in Eq. (7.5), $\alpha_0 = 0$ and $\alpha_1 = 1$, this would mean that subject i expects from the others that they will contribute in $t+1$ the same amount that they contributed in t; i.e., that the foundations of i's expectations are the actual contributions of his fellow members. I shall refer to the subjects for whom the latter holds as *perfect beliefs updating* (henceforward PBU) individuals.[22]

In a parallel to the analysis carried out in the previous subsection, before proceeding with the estimation of the regression coefficients for each individual and for each phase, I test for the equality of these coefficients in the two phases based on the same (basic) game. Specifically, for each subject, I estimate a model which uses the observations of both phases and which includes a dummy variable taking value 0 in the first phase and 1 in the third phase. If, in a regression, the difference in the coefficients is found to be insignificant, then I

[22] This terminology may appear somehow ambiguous. It intends only to indicate a person for whom there is a one to one correspondence between his beliefs in t and what he observes in t.

can pool the data over the two phases and have, regarding that regression, 18 (rather than 9) observations.

Results of this study reveal that, under the NCT, the difference between phases is significant for 4 people while, under the CT, it is significant for 8 subjects.

Let us start the analysis by considering the NCT data. Under this treatment, 5 out of the 20 subjects whose data can be pooled (i.e., 25%) exhibit a regression line with $\alpha_0 = 0$ and $\alpha_1 = 1$. As for the four subjects for whom the regression must be estimated separately for each phase, we find that three of them are PBU individuals in phase 3. By turning to the analysis of the relationship in the second phase of the NCT, the OLS regression gives a perfect fit in 4 cases (20%).

Under the CT, 6 out of the 16 pooled regressions (37.5%) present a constant equal to zero and a slope equal to one. Among the six subjects who show a statistically significant difference in their regression coefficients between phases, we observe one PBU individual in phase 1 and three PBU individuals in phase 3. Finally, as far as phase 2 is concerned, under the CT, the regression gives a perfect fit in 7 cases (29.2%).

Similar findings suggest that a few people in my experiment form and update their expectations on the basis of what they observe. This confirms previous experimental findings that the outcome of the game may affect the beliefs and the decisions in the next play of the game.[23]

7.7 Conclusions

The main purpose of this chapter was to test the relevance of the persuasive behaviour hypothesis in an environment more complicated than that used in the previous chapter. I did so by using a game with interior solutions (which, according to various authors, introduce scope for potentially confounding effects) and by exploiting the fact that, in this setting, different theories make different predictions about the relationship between a subject's contributions and the contributions of his partners.

Two main results come out from this study. The first is that (in contrast with the results obtained in the previous chapter) the persuasion hypothesis fails to explain much of this experiment's data: two subjects are categorised as persuaders under the CT and only one persuasive player is detected in each of the other two treatments.

[23] See, for example, Offerman (1996) and references quoted there.

The second noteworthy result of this experiment is that most of its data remain unexplained, which implies that also the other theories under investigation do not fare very well in this experiment. None of these earlier approaches to cooperation is able to account for the change in the same subject's behaviour that we observe within a same supergame.

In both the phases constructed around the basic game and under each of the three treatments considered here, we find a few subjects who contribute, unreciprocated, the socially optimal amount until their cutoff period, θ, and reciprocate the others' contributions thereafter. The regression that best tracks these subjects' behaviour turns, therefore, from zero to positive in period θ. While, in principle, the persuasive behaviour hypothesis can justify this kind of split, the restrictions imposed on persuasive behaviour do not allow the classification of a subject as a persuader if he is found to contribute the social optimum for more than one period. In other words, the 'hardwired' component of the persuasive strategy does not allow attempts to persuade to explain most observed behavioural patterns. Thus, it appears necessary to loosen these 'hardwired' elements of the program defining the basic persuasive strategy—by giving the agent increasing scope for discretion/optimisation—in order to improve the effectiveness of the persuasion hypothesis in explaining previously unexplained behaviour. But this is a different story which can only give suggestions for future improvements of the theory. For the moment, by relying on the definition adopted here and exposed in Chapter 5, we must reject the possibility of persuasion to be a valid explanation for the cooperative behaviour which people are found to exhibit in this (non–linear) experiment.

Chapter 8

Summary and suggestions for future research

8.1 Introduction

This concluding chapter attempts to provide an appraisal—based on the evidence from the experiments described in Chapters 6 and 7—of the results achieved so far by my behavioural hypothesis, an assessment of its strengths and weaknesses, and a vision of what might be done in the future.

Despite the relative robustness of the phenomenon of voluntary contributions in public goods environments, there is no unified theory which can explain all (or even most of) the observed regularities. So I begin my assessment by taking into account which kinds of previously unexplained empirical evidence the persuasion theory allows us to accommodate, underlying its main theoretical strengths (its simplicity and its resting on strong common–sense principles) and its main weaknesses (its limitations on the agent's discretion/rationality and its incapability of accounting for some of the former results). Then, I shall concentrate on the empirical validity of my theory by summarising how well it performed in the series of experiments which I have run to empirically examine it. Finally, I shall look to the future, indicating what developments should be done to reach more decisive conclusions.

8.2 Theoretical considerations

Let me recall the main motivations of this work: proposing a new theory of cooperative behaviour, and presenting a series of experiments which explore its empirical validity and, hence, which demonstrate if and to what extent it can be regarded as a positive theory of behaviour, where the latter refers to a theory

which is constructed as much as possible on empirically supported elements.[1]

The theory is relatively simple. It rests on two principles that seem to have strong common–sense. An agent understands the benefits that *he himself* can obtain if the public good is provided and, as a consequence, he contributes a constant amount for μ consecutive periods in order to persuade his 'selfish' partners to perform the same action. Then, if he is successful in these attempts, he continues with his constant contribution as long as he observes that each of the others does the same; otherwise, he modifies his contribution in the direction of the others' average contribution in the previous period, which implies that he decreases his contribution if it was above the average of the others. Hence, the persuasion theory is compatible with both the observed successes and the observed failures of voluntary cooperation.

It also allows for the accommodation of previously unexplained empirical evidence. In particular, two key observations from former dilemma games experiments are 'captured' by my theory. First, the frequency with which an individual chooses to cooperate after several cooperative outcomes have occurred; such a frequency implies that the individual does not take advantage of the others' willingness to cooperate by switching to the immediately rewarding (defecting) strategy. This seems to support a theory of reciprocity (according to which an individual's regard for the utility of the others depends on how 'kind' the others are towards him). But, if reciprocity is the principle underlying behaviour, it is not easy to explain a second key observation; namely, that people tend to repeat cooperative choices after they have just cooperated *without reciprocation*; experimental evidence shows, indeed, that there exist subjects who cooperate even if their group members do not do the same. Although pure altruism might account for this result, in contrast with altruism models, previous experimental findings report that cooperative subjects change their attitude towards the others if they refuse to modify their selfish behaviour.[2] An hypothesis as that proposed here—under which, a subject cooperates unconditionally for μ periods by expecting that, in $t = \mu$, the others will follow his behaviour, and then he reciprocates the others' observed decisions (so that he will stop cooperation if his expectations are not fulfilled)—appears to fit the reported observations quite well.

The fundamental assumption upon which I built my hypothesis is that, at the outset of the game, a persuasive player must believe that, if he contributes a constant amount throughout the first μ periods, all his partners will modify

[1] Cf., Wong (1987); and Offerman (1996).

[2] For discussion and experimental evidence related to this tendency, see Brewer and Kramer (1986); Dawes and Thaler (1988); Simon (1993); and Offerman (1996).

their initial 'selfish' choice and will follow his contributing decision from period μ until the end of the game. On the basis of such expectations he decides the optimal amount to contribute up to period μ. This is the only kind of discretion/rationality allowed to him: during the game, in fact, a persuasive player is a simple *automaton* forced to follow a predefined set of instructions.

I do not want to withhold the fact that a similar way of modelling cooperation has some inbuilt problems which may raise criticism. First of all, it restricts too much the rationality of an agent, who, once solved his first period's maximisation problem and decided (on the basis of a well–defined set of *a priori* expectations) how much to contribute throughout the first μ periods, must stick with the program of instructions given to him and cannot anymore optimise. This implies that, even if his *a priori* expectations change when the game is actually played, this change has no bearing on his 'automatic' behaviour: the set of instructions defining a persuasive strategy do not allow for a modification of the behaviour based on changed expectations. These kinds of restrictions bring the persuasive behaviour hypothesis in the line of earlier approaches to cooperation: a reciprocator, an altruist, a tit–for–tat player type can all be regarded as 'hardwired' individuals without any discretion during the game.

A further criticism which might be moved against my hypothesis is that it is incapable of accounting for some former experimental findings. Indeed, since persuasive motives can be pursued only in *repeated* public good games where players interact always in the *same groups*, the persuasion theory cannot explain why people cooperate in single–shot games or in repeated games where groups are randomly formed anew in each round. My response to a similar argument relates to the intentions I had in proposing a new model of cooperative behaviour. As pointed out earlier, I did not want to build one grand theory capable of explaining all significant phenomena of the games we are interested in. Such a theory seemed to me to be quite awkward and unmanageable. Rather, I wanted to propose a simple theory that could justify some previous unexplained empirical evidence.

8.3 Achievements to date

In principle, the hypothesis of persuasive behaviour deserves to be taken seriously as one among several ways to explain individuals' underlying motivations for choosing actions that do not maximise their monetary payoffs (especially in early stages of repeated social dilemma games).

In practice, since it generates testable predictions which differ from those of earlier theories, its ability to explain previously inexplicable behaviour can be empirically verified.

With the experiments described in Chapters 6 and 7, I wished exactly to assess the successes of my theory in comparison with alternative models of cooperative behaviour. This has been done in two ways. On the one hand, by exploiting the fact that the various theories make different predictions about the relationship between a player's contributions and the contributions of his group members; on the other hand, by adopting appropriate experimental designs. In particular, two aspects of my experiments were designed for serving this purpose. The first concerns the introduction (in both my experiments) of a questionnaire through which I asked subjects to motivate their own decision, to speculate about the motivation of the others and to predict the others' next decision. The second refers to the inclusion (only in Chapter 7's experiment) of a treatment in which subjects had to share the benefits deriving from the public good with a charitable organisation. Such a treatment allows for the verification of whether (or not) people support public goods when they do not derive private benefits from their production. The persuasive behaviour hypothesis—at odds with other theories—predicts that they should not.

8.3.1 On the strengths and weaknesses of my design

The inclusion of a questionnaire combining behaviour with beliefs elicitation and with a systematic investigation into subjects' motivations for their own decisions represents a very useful and straightforward means of obtaining insights into subjects' decision rules and better understanding the relationship between people's actions in public goods settings. Thus, in principle, its use can be considered a strength of my design. Nevertheless, the way in which I have implemented the questionnaire seems (in the light of the experience which I have gained in these kinds of experiments) not to be the best in order to identify a persuasive player–type.

Two obvious problems of my experiment are that it lacks any evidence on persuasive intention as well as on players' *a priori* expectations. Both these aspects are crucial for successfully detecting a persuader. On p. 3 of this work, I defined persuasion as the *intention* to influence the others in order to make them behave in a way that they otherwise would not. Furthermore, a persuasive player is predicted to hold a specific and well–defined set of *a priori* expectations about his partners' behaviour. This implies that if, before the start of the game, a subject is not observed to hold the kinds of expectations required by my

hypothesis, he cannot be classified as a persuasive type. Thus, the elicitation of each subject's *a priori* expectations would add a further testable restriction on persuasive behaviour and would contribute to an effective and unmistakable identification of this player type.

To reach the latter objective and correct the two weaknesses of my design above–mentioned, it seems opportune to amend the questionnaire in two of its features.

First of all, in order to know if a subject's contributing decision is intended to induce the others (who are free–riding) to do what is best for the group, it suffices to modify the list of alternatives among which the subjects have to choose in order to motivate their own decisions so as to include an answer explicitly corresponding to persuasion (e.g., "I want to persuade the others").

Second, in order to have information about the *a priori* beliefs of a player, it seems enough to elicit expectations before decisions. Although this modification would complicate the structure of each decision round,[3] it would permit the identification of the expectations that a subject holds prior to the start of the game so as to test if the prescriptions of the persuasive behaviour hypothesis about this issue are respected.

The latter modification would also bring my experiment closer to those run by Croson (1998a, 1998b), thus making a direct comparison between them easier. Since in my experiment, in contrast with Croson's ones, people are found to be more cooperative when the questionnaire is included, it might be argued that my opposite result is due to the fact that I applied the questionnaire after (rather than before) each decision. By carrying out an experiment that eliminates this source of difference, I can verify which between mine and Croson's findings are replicated. If the latter is observed, then we can infer that the moment in which beliefs are elicited affects behaviour in the sense that people react differently depending on whether their expectations are elicited before or after their decisions. Notice that this would simply be a 'framing' effect.

8.3.2 Major results: a summary

The main finding that comes out from my empirical analysis is that, in contrast with the predictions of economic theory, people voluntarily contribute to public goods. This is observed in both the experiments described in this work, which therefore are in line with many laboratory studies done by others. Data from

[3] As pointed on p. 119, since the motivational questions can be answered only after the decisions, the elicitation of the expectations before them would divide each single decision round into three (rather than two) parts. This is considered a complication of the design both for the subjects and for the experimenter.

Chapter 7's experiment reveal, however, that the willingness to contribute is significantly less when subjects do not get all the pecuniary benefits generated by the public good. Results from the latter experiment show, indeed, significantly differences in aggregate and individual behaviour between the supergames in which the subjects receive all the benefits associated with the production of the public good and the supergame in which they do not: individuals are found to free–ride significantly more in the latter than in the former. This suggests that the emphasis placed by the literature on people's 'unconditional' (i.e., without reserve) voluntary contributions may be misplaced: people appear willing to invest their own tokens in the production of public goods from which they (themselves) derive benefits, but not in the production of others.

A second noteworthy result (common again to both my experiments) is that a substantial number of subjects are found to modify their behaviour and their attitude towards the others within a supergame. This confirms the results of previous experiments and suggests that all theories which do not allow for this kind of modification cannot adequately explain much of my experimental data. Let us consider in detail how well the rival theories investigated in this work fared in my experiments.

A first hypothesis of cooperative behaviour which I examined is altruism. According to altruism theories, a player's utility increases not only in his own payoffs but also in the other players' payoffs. In Chapter 6's experiment, this implies unconditional cooperation whenever γ (the altruist's weight on the others' welfare) is greater than a critical value (which depends on the mrs between private and public goods). As for Chapter 7's experiment, in the two 10–fold repetitions of the basic game included in it, the assumptions of altruism theories imply unconditional cooperation and a negative reaction function for any value of γ; while, in the 10–fold repetition of the separation game, these implications are obtained only for values of γ greater than 0.7. Data from both the experiments reveal that a theory based on altruistic motives for giving is not adequate for explaining the cooperative behaviour exhibited by most of my subjects: participants in my experiments are found to decrease their contribution level if the other group members refuse to modify their 'selfish' attitude.

For the same reason motivations lying on Kantian principles are rather unsuccessful in explaining much of my data. If subjects were following a Kantian reasoning they should contribute the socially optimal amount throughout any supergame, regardless of the others' behaviour. But this is barely observed in both my experiments.

An alternative approach investigated in each of my experiments supposes

that people act on the basis of a reciprocity principle. According to such a principle, anyone who benefits from the public good has moral obligations towards those who contribute. Applied to my public goods situations, this implies that a subject must cooperate if he observes cooperation from all the others, and that he must increase his contribution if the others increase their contributions. Only 1% of the participants in Chapter 6's experiment are found to behave in accordance with the criteria used in that context to identify a reciprocator. Although reciprocity theories do much better in Chapter 7's experiment, they are unable to explain the kind of behavioural change that many subjects exhibit within a same supergame.

An alternative explanation as to why people cooperate is Andreoni's (1988) strategies hypothesis, which is a theory of rational behaviour in the sense of Kreps et al. (1982). According to it, when the information about the other types is incomplete, if a fully–rational and selfish player believes that there is a small chance that the others are irrational (i.e., non–payoff maximising), it may be rational for him to contribute in order to build reputation. This would imply a relatively high contribution level in the early periods, which decreases as the end draws near, given that, in the last round, the free–riding strategy is always optimal.

Although it would have been desirable to have the same set of theories open to investigation in each of the two empirical chapters, the impossibility of deriving the form of a strategic player's reaction function in the games around which Chapter 7's experiment was constructed led me to not include 'strategies' among the theories investigated in the latter chapter. As far as the data from Chapter 6's experiment are concerned, strategic behaviour is rather successful in justifying them: 21 percent of the participants in this experiment are found to behave according to the criteria used in that context to identify a strategic player; such a percentage decreases to 14% if (in analysing the data relative to the second subsession)[4] we see phases 4, 5 and 6 as continuations of phases 1, 2 and 3 respectively. Specifically, it is the observation of some subjects who cooperate throughout a supergame (in the sense that they are found to not defect in the final period) which weakens this approach's ability to explain my data.

Finally, let us consider how well the hypothesis that I propose in this work to explain why people voluntarily provide public goods performed in each of my two experiments. The results obtained in this respect are contrasting: while the data from my first 'simple' experiment show that a quite reasonable per-

[4]Recall that the experiment described in Chapter 6 consisted of two perfectly identical subsessions.

centage of the observed cooperative behaviour—which cannot be explained by any of the other alternative theories—is explicable with the persuasion hypothesis,[5] the latter fails to explain much of the data from the more complicated (non–linear) settings upon which Chapter 7's experiment was built.[6] Nevertheless, the observation, in this experiment, of a few subjects who contribute, unreciprocated, the socially optimal amount for some consecutive periods and reciprocate the others' behaviour thereafter suggests that it is because of the restrictions imposed on the rationality of a persuasive player that attempts to persuade are unable to explain a higher percentage of the observed behavioural patterns.[7] Thus, it appears necessary to loosen the 'hardwired' elements of the program defining the basic persuasive strategy—by giving the agent increasing scope for discretion/optimisation—in order to improve the effectiveness of the persuasion hypothesis in explaining previously unexplained behaviour.

8.4 Other theories and other empirical studies

Other approaches in the literature might be relevant for describing my experimental results. One of these is the equity (or inequality aversion) theory proposed by Fehr and Schmidt (1999), and Bolton and Ockenfels (2000). According to equity theories, a subject's regard for the others' payoffs depends on how the latter compare to his own. These models are based on the assumption that people dislike inequality in payoffs, and that they dislike inequality more if it is to their disadvantage than if it is to their advantage. This implies a contributing behaviour which depends on fixed (exogenous) preferences over payoff distributions, regardless of whether the others have done anything at all. Nevertheless, players have to form beliefs about the others' choices in order to choose their contribution level and, as long as inequality–averse players believe that other players are contributing, they are willing to contribute too. If (as found in my experiments) subjects update their beliefs in the light of what they observed in the previous round, a few patterns of behaviour detected in my experiments are compatible with equity (inequality aversion) theories: if matched with other contributors, contributive subjects keep on contributing throughout the game; if matched with defectors, they lower their own contribution. These

[5]Specifically, if we consider the two identical subsessions of this experiment as independent, we found that, out of a total of 144 individuals (24 subjects × 6 phases), 14 percent act in accordance with my hypothesis of persuasive behaviour. This percentage increases to 20% if we consider the second subsession as a continuation of the first one.

[6]Only 3 out of the 72 participants in this experiment are, in fact, categorised as persuaders.

[7]Recall that, in Chapter 7's experiment, a subject cannot be classified as a persuader if he is found to contribute the social optimum for more than one period.

theories cannot, however, explain the fact that some contributive subjects are found to repeat their choice for some consecutive periods before matching the others' lower contributions. Even if their fixed preferences could be such that they solve the trade–off between pecuniary and relative payoffs in favour of the latter (so as to justify the repeated, unreciprocated, early contributions), it is not clear why these preferences should change over time.[8]

On the other hand, there is some evidence from recent experiments which corroborates my findings that subjects unconditionally cooperate in early periods of repeated public goods games, and then they reciprocate the observed contributions of their partners. This is reported, for instance, by Keser (1999) and by Keser and van Winden (2000), who suggest an interpretation of subjects' behaviour in terms of (what they call) *conditional cooperation*, which is based on the two aspects of future–oriented and simple reactive behaviour. As future–oriented behaviour they define "aspects of subjects' behaviour that are induced by their perception of future interaction"; as reactive behaviour they intend individuals' behaviour that changes in the direction of the other group members' average contribution in the previous period.

If we interpret my experimental results as suggesting the necessity to combine the insights of various approaches into a richer model in order to adequately explain the voluntary contributions phenomenon, we find further support in the existing literature, such as in Bolton (1998) and Charness and Haruvy (1999).

8.5 Research agenda

In looking towards the future, there seem to me to be two major issues that I will need to confront to assess the validity of my persuasive behaviour theory. The first is a purely theoretical issue and concerns the (already emphasised) necessity to loosen the 'hardwired' elements of the program defining the basic persuasive strategy, giving the agent more discretion with respect to the length of his reciprocity–test–period. This might be done by allowing the persuasive player to Bayesian update his expectations during the game, rather than requiring him to stick with the optimal contribution calculated at the outset of the game on the basis of his *a priori* expectations.

The second issue relates to my methods which, as pointed out in Subsection (8.3.1), should be amended with respect to the implementation of the questionnaire.

[8] A fundamental assumption of Bolton and Ockenfels (2000) is that the trade–off must remain stable for the duration of the experiment.

In the immediate future, I intend to replicate the 'simple' experiment described in Chapter 6 whose main problem, besides the two just mentioned, is the small number of observations. Indeed, the experiment consisted of only two sessions with 12 subjects each. Since public goods decisions are of high variance and earlier research has shown that people differ very much in their inclination to cooperate, my strong interest in individual characteristics requires much more data than those collected in the study presented in this work. More data would increase the robustness of the classifications and, hence, the confidence in my findings.

In replicating such an experiment, apart from modifying the definition of the (basic) persuasive strategy and the way in which I conducted the questionnaire, I intend to add two control treatments. The first is "borrowed" from my second experiment, and refers to the running of an experimental session without the questionnaire in order to check if and in which way the questions/predictions asked between rounds affect the rate of cooperation.

The other control treatment which I would like to run refers to the possibility of transforming a theoretical weakness of my model (i.e., its impossibility of accounting for some former results) into an empirical strength. This would be done by taking into account (along with Partners) Strangers situations. Specifically, a same pool of subjects would be required to participate in two different sessions: in one (the Partners session), they would interact in the same groups during the entire game (like the experiments presented here); in the other (the Strangers session), the groups' composition would change in each period in a way such that participants cannot expect to meet the same fellow members again in a later period. In order to control for *order effects*, I would carry out two treatments which differ from each other depending on having the Partners session before or after the Strangers session. A similar experimental design would allow me to verify if cooperative agents classified as persuasive in the Partners situation change their behaviour (and, therefore, defect) in the Strangers situation. I can, in such a way, reach more decisive conclusions about the practical validity of the persuasive behaviour hypothesis.

8.6 Conclusions

The hypothesis that, in social dilemmas, a subject cooperates because of his understanding of the benefits that he derives from such a choice when it is made by other group members, and of the consequent benefits from him to induce the others to cooperate, finds some support in one of the experiments

reported in this work, but it fails to explain much of the data from a second (more complicated) experiment. In this concluding chapter, I have proposed a number of improvements which can be made both to the theory and to the methods used for testing it so as to reach a better assessment of its ability to explain the observed voluntary contributions phenomenon.

Appendix A

First experiment's instructions

This appendix contains the instructions given to the participants in the experiment described in Chapter 6. Section A.1 provides the instructions for the first subsession. Section A.2 presents those for the second subsession.

A.1 Instructions for the first subsession

This experiment is a study of economic decision making. The instructions are simple. If you follow them carefully and make good decisions, you may earn a reasonable amount of money.

The unit of EXPERIMENTAL MONEY will be the TOKEN; i.e., during the experiment you will be earning tokens. These will be converted into pounds and paid in cash to you at the end of the experiment according to the exchange rate 1 Token = £0.02.

The University of York has provided the funds for this study.

Please take your time to read these instructions at your own pace. If you have any questions while reading them, please raise your hand and someone will come to help you.

THE EXPERIMENT

The experiment consists of 2 PARTS. There will be a ten–minute break in the middle. You will receive the instructions for the first part in a moment. Information on the second part will be given to you after the break.

DESCRIPTION OF THE FIRST PART

The first part of the experiment is divided into 3 PHASES of 10 ROUNDS each. This means that you will play for a total of 30 rounds.

Groups At the beginning of each phase, you will be assigned to a group of 3 people.

The composition of your group will remain FIXED throughout each phase. That is, you will play with the same other two subjects in all 10 rounds that each phase contains.

The composition of your group will be changing every phase. After each phase you will be REASSIGNED TO A NEW GROUP of 3 participants. The 3 group members will never have been members of the same group in the past. You have NO chance of being in a group with any other participant more than once.

At no point in the experiment will you know the identities of the other 2 members of your group, nor will your identity be made known to them.

You are not allowed to talk or communicate with other participants.

The One–Round Choice In each round of part 1, you and the other two members of your group will EACH have two possible choices.

YOU MUST CHOOSE BETWEEN **X** and **Y** without knowing what the others in your group are deciding. The tokens you get will depend on the choice that you and the other members of your group make.

1. **Choice of X**

If you choose X, you (and ONLY YOU) will earn 50 tokens. The other members of your group will receive NOTHING as a consequence of your choice.

2. **Choice of Y**

If you choose Y, you, as well as EACH MEMBER OF YOUR GROUP, will earn v tokens. Similarly, YOU (and everyone in your group) will receive v tokens for each of your group–members who choose Y. That is, it does not matter who chooses Y: every member of the group gets v tokens from it—whether he chose Y or not.

We will reveal to you what v is during the experiment.

The Questionnaire In every round, besides making a choice between X and Y, you will be also asked:

1) to indicate the main reason for your choice in that period;

2) to guess what motivated the previous choice of both your fellow–participants;

3) to predict the next choice of both your fellow–participants.

A questionnaire will appear on your computer screen after all three group members have made their choices asking you to pick out one (and ONLY ONE) of several alternatives for every question.

Additional Tokens The questionnaire will allow you to get tokens in addition to those you accumulate from your choices. In fact, your predictions of your fellow–participants' next choice as well as your beliefs about the motivations behind their last choices will be confronted with their answers.

Each round you may earn 30 ADDITIONAL TOKENS if you predicted correctly the next choice of BOTH your partners. A FURTHER 30 TOKENS will be given to you if you guessed the motivation behind the last choices of BOTH your fellow–participants.

We stress that we do NOT consider your answer correct if you guessed the next decision and the motivation of ONLY ONE member of your group.

Time Limits You can take your time to give your responses in the first round of each phase. But in all other 9 rounds that each phase contains you will have TIME LIMITS on each of your responses. This time is fixed to 20 SECONDS PER ANSWER, so that you have to take each decision in 20 seconds.

A countdown timer displaying the number of seconds remaining in the current decision will be visible on your screen and 5 seconds before the end of the time you will hear a sound informing you that you have to take a decision very quickly. If you answer after the time limit, your response will be taken to be that of the previous round.

Recalling Previous Choices Before starting a new round you may want to recall your partners' previous decisions. By pressing **P**—at the end of each round—you will be able to see the past history of the game.

Your Cash Earnings The tokens you obtain on all 3 phases of the first part will be ADDED to calculate your experimental earnings from that part.

This experimental earnings will be then turned into cash earnings. In particular, your cash earnings from the first part will be the total number of tokens you earned in it divided by 5. For example, if you get 4000 tokens, your cash earnings will be 800 p. If you get 3000 tokens, then your cash earnings will be 600 p.

INFORMATION ON THE SCREEN

Payoff–Table In the figure which follows is depicted the PAYOFF–TABLE PER ROUND.

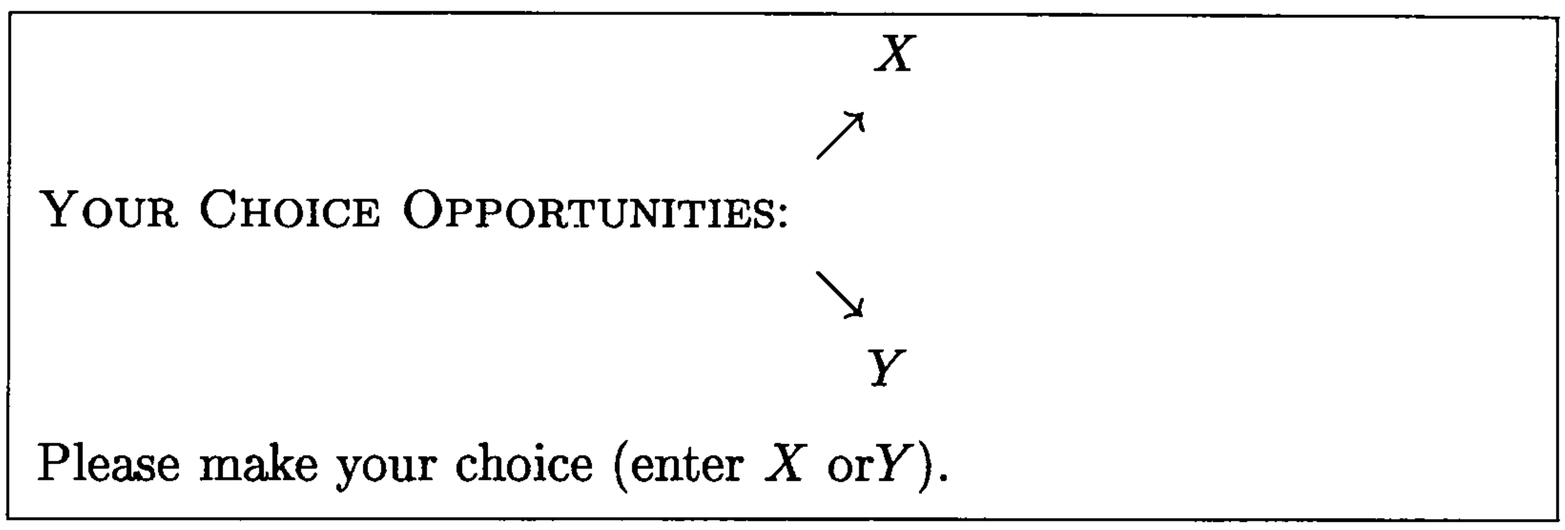
Depending on your choice, and everyone else's in your group, six different situations may occur in each round. In particular, if you choose **X**, you may

	(a) r	if all three group members choose X
/	(b) $r+v$	if one person in the group chooses Y
X/	(c) $r+2v$	if both the other two fellow participants choose Y
/		
\		
Y \	(d) $3v$	if all three group members choose Y
\	(e) $2v$	if one person in the group chooses X
	(f) v	if both the other two fellow participants choose X

obtain one of the situations indicated by (a), (b) and (c). If you, instead, choose **Y**, one of the situations indicated by (d), (e) and (f) may occur.

The exact value of v will be revealed to you during the experiment.

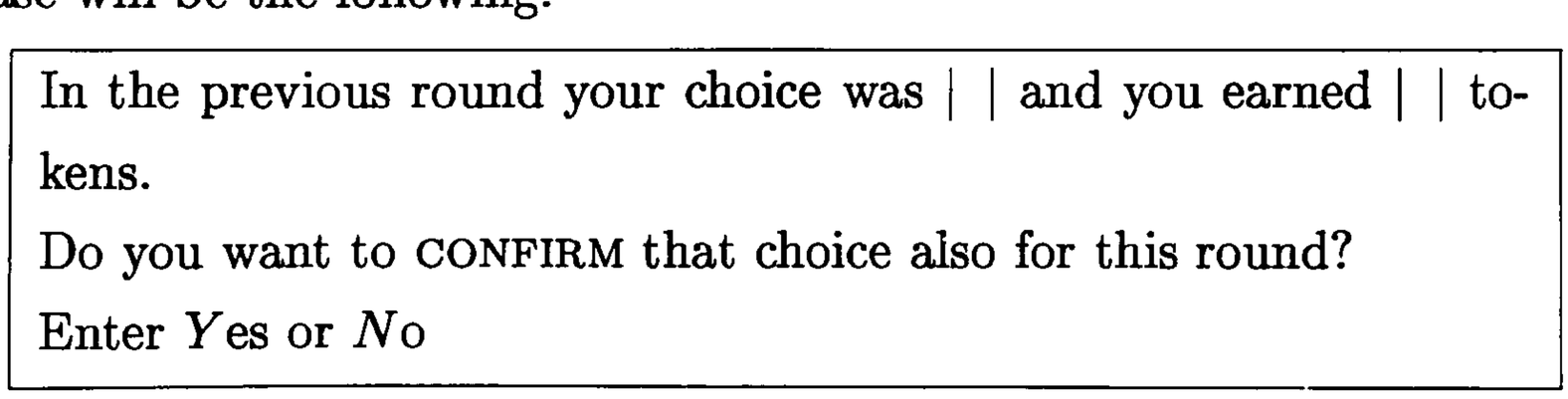
The Decision–Window for Making a Choice To report your choice you will have to use a 'decision window' which in the FIRST ROUND of each phase will be like the one shown below.



In this window, your two choice opportunities are shown and by typing X or Y you can indicate for which opportunity you want to enter your decision.

After you enter a choice, the three possible situations related to that choice will appear on your screen. If you are NOT happy with your choice, you have the possibility to change it by pressing N.

The Decision–Window for Confirming your Choice The decision window that will appear on your computer screen in ALL OTHER ROUNDS of each phase will be the following:



In this window, both the action you chose and the tokens you earned in the previous round are shown. So that, if you are still happy with that choice you can simply confirm it and continue with the experiment by pressing Y.

Otherwise, if you are not more happy with your previous choice, you can modify it by typing N.

Result–Table

RESULTS TABLE FOR ROUND NUMBER: q
YOUR choice:
Choice of the SECOND MEMBER of your group:
Choice of the THIRD MEMBER of your group:
Situation occurred: g ($g = a,b,c$ if you chose X) ($g = d,e,f$ if you chose Y)
Your Payoff
TOTAL NUMBER OF TOKENS YOU GOT FROM YOUR CHOICES:

When everybody in your group has made his choice, a RESULTS TABLE (shown on the left) will appear on your screen.

In this table, your choice, the choices of your fellow participants, and the situation occurred in that round (with your related payoff) are registered.

The total number of tokens you earned from your choices in all past rounds is shown at the bottom of the table.

The Questionnaire After having seen the results table, you will have to answer three questions.

The FIRST QUESTION asks you to indicate the main reason for your choice in that round. In order to answer it, you must choose one (and ONLY ONE) of the 5 alternatives listed in Table A.1.

INDICATE THE 'MAIN' REASON FOR YOUR CHOICE:
I chose | | because:
A) of the past behaviour of my fellow-participants;
B) whatever my fellow-participants would have chosen, this choice assured me higher payments than those I could obtain by choosing the other action;
C) if all 3 group members took this choice, we would obtain the highest payments;
D) this choice gave my fellow-participants the lowest payments;
E) of other reasons. Please, specify which these reasons are.

Table A.1: First question of the questionnaire

Once you have taken your decision, you must type the associate letter on your keyboard.

The SECOND QUESTION asks you to guess what motivated the previous de-

cisions of your fellow–participants. You must answer this question twice (once for the second member of your group and once for the third member) and each time you must pick out one (and ONLY ONE) of the 5 alternatives listed in Table A.2.

Why, in your opinion, did the SECOND (THIRD) MEMBER of your group choose as he (or she) did?
A) because of my and third member's past behaviour;
B) because whatever me and the third member would have chosen, this choice assured him (her) higher payments than those (s)he could obtain by choosing the other action;
C) because he realised that, if all three group members took this choice, we would obtain the highest payments;
D) because that choice gave other participants the lowest payments;
E) because of other reasons.

Table A.2: Second question of the questionnaire

The THIRD and LAST QUESTION asks you to predict the next choice of each of your fellow–participants.

Once you have made your predictions, you must type the choice you think your fellow–participants will take (i.e., X or Y) on your keyboard.

A.2 Instructions for the second subsession

The second part of the experiment will be an EXACT REPETITION of the first part.

This means the following.

1) You will play 3 different PHASES of 10 ROUNDS each in a group of 3 people.

2) The 2 participants you will be matched with are THE SAME as in the first part; i.e., you will play each ten–round phase of the second part with the same people with whom you played in the respective phase of the first part.

3) In each round you will be asked to choose between X and Y, and to answer the questionnaire.

4) The 3 values of v will be exactly like they were in the first part.

5) Your cash earnings will be calculated as in the previous part.

Appendix B

Second experiment's instructions

This appendix contains the instructions given to the participants in the experiment described in Chapter 7. Specifically, it presents the instructions for the communication treatment.

Omitting *either* the text in italics on page 206 *or* all the references to the questionnaire, one obtains the set of instructions which were provided in the non–communication treatment and in the non–questionnaire treatment, respectively.

Instructions for the communication treatment

This experiment is a study of economic decision making. The instructions are simple. If you follow them carefully and make good decisions, you may earn a reasonable amount of money.

The unit of EXPERIMENTAL MONEY will be the FRANC; i.e., during the experiment you will be earning francs. These will be converted into pounds and paid in cash to you at the end of the experiment according to the exchange rate 1 Franc = £0.50.

In a way that will be explained later, through your decisions you may also help a charity organization. Next to your computer, there is a list of charities. After you have completed the experiment, you will be asked to choose one of them—if you decided to help any.

The European Commission has provided the funds for this study.

THE EXPERIMENT

The experiment is divided into 3 PHASES of 10 ROUNDS each. You will be therefore playing for a total of 30 rounds.

Groups At the beginning of each phase, you will be assigned to a group of 3 people.

The composition of your group will remain FIXED throughout each phase. That is, you will play with the same other two subjects in all 10 rounds that each phase contains. After each phase you will be RANDOMLY REASSIGNED TO A NEW GROUP of 3 participants.

The 3 group members will never have been members of the same group in the past. You have NO chance of being in a group with any other participant more than once.

At no point in the experiment will you know the identities of the other 2 people in your group, nor will your identity be made known to them.

You are not allowed to talk with the other participants.

The One–Round Investment Decision In each round of the experiment, you and the other 2 members of your group will EACH have a budget of 9 tokens.

You must decide—without knowing what the others in your group are deciding—how many of these tokens you wish to invest in a public account.

Each decision you take yields a given amount of francs to YOU and to EACH MEMBER OF YOUR GROUP. The number of francs you may earn depends therefore on the investment decisions of ALL group members; it is reported in a PAYOFF MATRIX. This matrix will appear on your screen at the beginning of each round. It will be the same throughout all ten periods of each phase.

The Payoff Matrix in the First and Third Phases The matrix that you will face in the first and third phases is depicted in Table B.1. The ROWS of the matrix represent the number of tokens you decide to invest; the COLUMNS are the total number of tokens invested by both your partners, and its ENTRIES show the amount of francs you may earn according to your and your partners' investment decisions.

The Payoff Matrix in the Second Phase In the second phase, the numerical details of the payoff matrix will differ. The new matrix is described in Table B.2. The consequences of your decisions will be different as well: through them, you may provide money for a charitable institution.

Charitable Contributions Any token you decide to invest in phase 2 will generate earnings for a charity organization.

UNLESS YOU INVEST ZERO, you will keep only half of the entries in the matrix; the remaining half will be given to the charitable institution that you will pick out from the provided list.

ONLY IF YOU INVEST ZERO, you will receive the entire amount of francs written in the first row of the matrix.

Your earnings will depend—of course—on the decisions of the other two group members. Suppose, for example, that the total investment of your partners is 5 tokens; then, if you invest 3 tokens, your own earnings would be $\frac{46}{2} = 23$ francs, and the same amount of francs will be given to the charity; if, instead, you choose to invest 0, then your earnings would be 62 francs and the charity will receive nothing as a consequence of your decision.

The Questionnaire In every round, besides making an investment decision, you will be also asked:
1) to indicate the main reason for your decision in that period;
2) to guess what motivated the previous decision of both your fellow–participants;
3) to predict the next decision of both your fellow–participants.

A questionnaire will appear on your computer screen after all three group members have made their investment decisions asking you to pick out one (and ONLY ONE) of several alternatives for every question.

The reason you have selected for your investment decision will be said to your partners and you will be informed about your partners' actual motivations for their decision.

The FIRST QUESTION asks you to indicate the main reason for your own investment decision in that round. In order to answer it, you must choose one (and ONLY ONE) of the answers listed below:
1) I wanted to make MY OWN earnings as large as possible given that I was expecting the same behaviour from the others;
2) if ALL 3 group members decided to invest that amount of tokens, WE would obtain the highest payments;
3) I thought that the others would have invested a LARGE amount of tokens;
4) I thought that the others would have invested a SMALL amount of tokens;
5) I wanted to help my partners;
6) that investment decision gave my partners the lowest payments;
7) the others' past behaviour induced me to make that decision;
8) time was over;

9) IF PHASE 2: I wanted to help the charity.

If none of the provided alternatives expresses the right motivation for your decision, then—by typing **0**—you will have the chance to write down your specific reasons.

The SECOND QUESTION asks you to guess what motivated the previous investment decision of your fellow–participants. You must answer this question twice (once for each member of your group) and each time you must pick out one of the following alternatives:

0) he provided his own reasons;
1) he wanted to make HIS OWN earnings as large as possible given that he was expecting the same behaviour from the others;
2) he realised that—if ALL 3 group members decided to invest that amount of tokens—WE would obtain the highest payments;
3) he thought that the others would have invested a LARGE amount of tokens;
4) he thought that the others would have invested a SMALL amount of tokens;
5) he wanted to help the others;
6) his investment decision gave the partners the lowest payments;
7) the others' past behaviour induced him to take that decision;
8) time was over;
9) IF PHASE 2: he wanted to help the charity.

Once you have chosen one of the listed alternatives, you must enter the number associated with it.

The THIRD and LAST QUESTION asks you to predict how many of their 9 tokens will be invested by EACH of your partners in the following round.

Once you have made your predictions, you must type the expected amounts (i.e., two numbers from 0 to 9) on your keyboard.

Additional Francs The questionnaire will allow you to get francs in addition to those you accumulate from your investment decisions. In fact, your beliefs about the motivations behind your partners' last decision, as well as your predictions of their next decision, will be compared to their answers.

Each round you may earn 50 ADDITIONAL FRANCS if you guessed correctly the motivation behind the last decision of BOTH your fellow–participants.

A FURTHER 50 FRANCS will be given to you if you predicted correctly the next decision of BOTH your partners.

We stress that we do NOT consider your answer correct if you guessed the motivation and the next decision of ONLY ONE member of your group.

Time Limits You can take your time to give your responses in the first round of each phase. But in all other 9 rounds that each phase contains you will have TIME LIMITS on each of your responses.

This time is fixed to 20 SECONDS PER ANSWER, so that you will have:

1. 20 seconds to decide how many tokens you wish to invest;
2. 20 seconds to provide the main reason for that investment decision;
3. 20 seconds to guess the motivation of BOTH your partners;
4. 20 seconds to predict BOTH their next investments.

A countdown timer displaying the number of seconds remaining in the current decision will be visible on your screen and 2 SECONDS before the end of the time you will hear a sound informing you that you have to take a decision very quickly. If you answer after the time limit, your response will be taken to be that of the previous round. So that, it does not make sense that you try to modify your decision after the permitted time.

Recalling Previous Choices Before starting a new round you may want to recall your partners' previous decisions. By pressing **P**—at the end of each round—you will be able to see the past history of the game.

Moreover, next to your computer, you can find a record sheet where you can write down—if you wish—the round by round decisions of your partners.

Your Cash Earnings The francs you obtain on all phases will be ADDED to calculate your total experimental earnings.

These experimental earnings will be then turned into cash earnings. In particular, your cash earnings will be the total amount of francs you earned divided by 2. For example, if you get 4000 francs, your cash earnings will be 2000 p. If you get 3000 francs, then your cash earnings will be 1500 p.

		Total investment by the other two group members																		
		0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
Your investment	*0*	16	23	30	36	41	46	51	55	59	63	67	71	74	78	82	86	88	92	95
	1	19	26	32	37	42	47	52	56	60	64	68	72	75	78	81	85	88	90	93
	2	22	27	33	38	44	48	51	55	59	63	66	70	73	76	79	82	85	88	91
	3	23	28	33	38	42	46	50	54	58	61	65	68	71	74	77	80	82	85	87
	4	24	28	33	37	41	45	49	52	56	59	62	65	68	71	74	76	79	81	84
	5	24	29	32	36	40	43	47	50	53	56	59	62	65	68	70	73	75	78	80
	6	22	27	30	34	38	41	44	47	50	53	56	59	61	64	66	69	71	73	76
	7	21	25	28	32	35	38	41	44	47	50	52	55	57	60	63	64	66	69	71
	8	18	22	25	28	32	34	37	40	43	45	48	50	52	55	57	59	61	63	65
	9	15	19	22	25	27	30	33	35	38	40	43	45	47	49	51	53	55	57	59

Table B.1: The payoff matrix in Phases 1 and 3.

		Total investment by the other two group members																		
		0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
Your investment	*0*	32	39	46	52	57	62	67	71	75	79	83	87	90	94	98	102	104	108	111
	1	19	26	32	37	42	47	52	56	60	64	68	72	75	78	81	85	88	90	93
	2	22	27	33	38	44	48	51	55	59	63	66	70	73	76	79	82	85	88	91
	3	23	28	33	38	42	46	50	54	58	61	65	68	71	74	77	80	82	85	87
	4	24	28	33	37	41	45	49	52	56	59	62	65	68	71	74	76	79	81	84
	5	24	29	32	36	40	43	47	50	53	56	59	62	65	68	70	73	75	78	80
	6	22	27	30	34	38	41	44	47	50	53	56	59	61	64	66	69	71	73	76
	7	21	25	28	32	35	38	41	44	47	50	52	55	57	60	63	64	66	69	71
	8	18	22	25	28	32	34	37	40	43	45	48	50	52	55	57	59	61	63	65
	9	15	19	22	25	27	30	33	35	38	40	43	45	47	49	51	53	55	57	59

Table B.2: The payoff matrix in Phase 2.

Appendix C

First experiment's data

This appendix reports all individual data collected from the experiment described in Chapter 6. Section C.1 describes the data from the first subsession. Section C.2 describes those from the second (identical) subsession. In both cases, data appear separated according to phase.

For all 10 periods of the 6 phases in which the experiment was divided, each of the tables that follow reports:

i) the choice between X and Y made by the subject in the decision stage, in the first column;

ii) the motivation provided by him for that choice, in the second column;[1]

iii) his guesses about the motivations underlying the second and the third partners' choices, in the third and forth columns respectively;[2]

iv) his predictions about the next choices of the second and the third partners, in the fifth and sixth columns respectively.

Throughout all exposition, the three aligned tables refer to the three players in the same group.

---

[1]Table A.1 in Appendix A depicts the 5 alternatives (labelled from A to E) among which subjects had to choose in order to answer this question.

Actually, Table A.1 shows the 5 possible answers available to the participants from the second to the tenth round of each phase of subsession 1, and in all ten rounds of each phase of subsession 2.

In the first round of each phase of subsession 1, alternative A was not in the list and the 4 remaining alternatives were labelled A, B, C, and D, respectively.

[2]Table A.2 in Appendix A lists the 5 alternatives among which subjects had to choose in order to answer this question. Also in this case, alternative A was not included in the first round of each phase of subsession 1.

C.1 First subsession's data

Phase 1

Subject 1

Y	B	A	B	X	X
Y	C	B	C	X	Y
Y	C	B	A	Y	Y
Y	C	B	A	X	Y
X	B	B	C	X	Y
X	B	B	C	X	Y
X	A	A	C	X	Y
X	A	A	C	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 4

X	A	B	B	X	X
X	B	C	C	Y	Y
X	B	C	A	X	Y
X	B	A	A	X	X
X	A	A	C	X	X
X	A	B	C	X	Y
X	B	A	C	X	Y
X	B	A	C	X	X
X	A	A	A	X	X
X	B	A	A		

Subject 9

Y	B	B	D	Y	Y
Y	C	A	D	X	X
Y	A	A	B	Y	Y
Y	E*	A	B	X	X
Y	C	A	B	Y	Y
Y	C	E	E	Y	Y
Y	C	E	E	Y	Y
Y	C	A	B	X	X
X	A	A	A	X	X
X	A	E	A		

* Hope others will follow.

Subject 2

X	A	B	B	Y	Y
X	A	C	C	Y	X
X	A	B	C	X	Y
X	A	A	A	X	Y
X	A	A	E	X	Y
X	A	A	E	X	Y
X	A	A	D	X	Y
X	A	A	D	X	X
Y	A	A	D	X	X
X	C	C	D		

Subject 5

Y	B	A	B	Y	Y
Y	C	B	C	X	X
X	A	B	C	X	X
X	A	B	C	X	X
X	A	B	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	E	A	X	X
X	B	B	B		

Subject 10

Y	D*	A	B	X	Y
Y	C	B	C	X	X
Y	C	B	A	Y	Y
Y	C	B	A	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	E	B	X	X
X	A	B	B		

* If we all chose this the group as a total would benefit

Subject 7

X	A	A	B	X	X
X	A	A	C	X	X
X	B	A	C	X	X
X	A	A	C	X	X
X	D	A	E	X	X
X	B	B	C	X	X
X	B	B	A	X	X
X	B	B	A	X	X
X	B	B	A	X	X
X	B	B	A		

Subject 8

X	B	B	C	X	X
X	A	B	C	X	X
X	A	B	C	X	X
X	B	B	C	X	X
X	A	B	C	Y	X
X	A	E	C	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 12

Y	B	A	A	X	X
Y	C	B	E	X	X
Y	E*	A	A	Y	X
Y	C	A	A	X	Y
Y	C	A	A	Y	Y
Y	C	A	A	X	X
X	A	B	B	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

* Expected at least another to follow.

Subject 3

X	A	A	A	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	A	A	A	X	X
Y	E*	A	A	X	Y
X	A	A	A	X	Y
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

* Make my partners to choose differently in the next rounds

Subject 6

X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 11

X	D	D	D	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	E	A	X	X
Y	E*	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

* Used it as a probe to see if others would choose Y.

Subject 13

X	A	A	A	Y	X
X	E*	E	E	X	X
X	A	B	E	X	Y
X	A	A	B	X	X
Y	B	B	B	Y	X
Y	A	A	A	X	X
X	B	A	A	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

* I felt that it was the best profit based on my predictions of their actions.

Subject 16

X	A	A	A	X	X
X	B	B	C	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	C	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 21

X	A	A	A	X	Y
X	B	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	C	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 14

Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	A	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	C	A	C	Y	Y
Y	C	B	A		

Subject 17

Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	Y
X	A	A	C		

Subject 22

Y	B	B	B	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	B		

Subject 15

X	A	D	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	E	E	X	X
X	A	C	E	X	X
Y	E*	C	E	Y	X
Y	E*	E	E	Y	X
Y	E*	E	E	Y	Y
Y	E*	E	E	X	X
X	A	E	E		

* If we all chose this we' d be better off

Subject 18

X	A	A	A	X	X
X	A	A	A	Y	Y
X	A	A	A	X	X
X	A	A	A	X	X
Y	C	A	A	Y	Y
Y	C	A	A	X	X
X	A	E	A	X	X
X	A	E	A	X	X
X	A	E	A	X	X
X	A	E	A		

Subject 23

X	A	A	A	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	C	X	X
X	B	C	C	X	Y
X	B	C	B	X	X
X	B	C	B	Y	X
X	B	C	B	X	X
X	B	B	B		

Subject 19

X	A	A	A	Y	Y
X	B	E	E	X	X
X	B	C	B	X	X
X	B	C	B	X	X
X	B	B	B	X	Y
Y	A	B	B	X	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	X
Y	A	A	B	Y	Y
Y	A	A	B		

Subject 20

X	A	A	A	Y	Y
Y	C	B	C	X	Y
Y	C	A	C	Y	Y
Y	C	A	C	X	X
X	A	B	B	X	X
X	A	A	B	Y	X
Y	B	A	A	Y	Y
Y	B	C	A	Y	X
Y	A	C	A	Y	Y
Y	C	C	B		

Subject 24

X	A	A	A	Y	X
X	B	A	A	Y	X
X	B	A	A	X	Y
X	A	A	A	X	Y
X	A	A	A	X	Y
X	A	A	A	Y	X
X	A	A	A	Y	X
X	B	A	A	Y	X
X	B	A	A	X	Y
X	B	A	A		

Phase 2

Subject 1

Y	B	A	A	X	X
Y	C	B	B	Y	X
Y	C	B	B	Y	Y
Y	C	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 11

X	A	D	D	X	X
X	B	C	B	X	X
X	B	C	B	X	X
X	B	C	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 5

X	A	C	A	X	X
X	A	C	A	X	X
X	B	C	A	X	X
X	B	C	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 2

Y	B	A	B	X	X
X	A	B	B	X	X
X	A	C	B	X	Y
X	A	C	C	X	X
X	B	C	B	Y	X
X	B	B	B	X	Y
X	B	B	C	X	Y
X	B	B	C	X	Y
X	B	B	B	X	X
X	B	B	B		

Subject 9

Y	B	D	B	Y	Y
X	A	A	D	X	X
X	A	C	D	Y	Y
Y	C	C	B	X	X
X	A	C	B	X	X
X	A	A	B	Y	X
Y	C	A	B	Y	X
Y	C	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 12

X	A	D	D	X	X
X	A	B	A	X	X
Y	C	B	A	Y	Y
Y	C	C	A	Y	X
Y	C	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	B	C	A	X	X
X	B	A	A	X	X
X	B	A	A		

Subject 3

Y	B	A	B	X	X
X	A	B	B	X	X
X	A	C	B	X	Y
X	A	C	C	X	X
X	B	C	B	Y	X
X	B	B	B	X	Y
X	B	B	C	X	Y
X	B	B	C	X	Y
X	B	B	B	X	X
X	B	B	B		

Subject 7

Y	B	A	B	X	X
Y	D	D	B	X	X
Y	B	E	B	X	Y
Y	C	C	C	X	Y
Y	B	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	B	C	C	Y	Y
Y	B	B	B		

Subject 10

Y	B	A	B	Y	Y
Y	C	B	C	X	X
X	A	A	C	Y	X
Y	A	A	C	Y	Y
Y	C	C	A	Y	Y
Y	C	C	A	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	B	C	C		

Subject 4

X	A	A	B	X	X
X	B	B	C	X	X
X	A	B	C	X	X
X	A	A	A	X	Y
Y	C	A	A	X	Y
Y	C	B	C	X	Y
Y	C	B	C	X	X
X	D	B	C	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 6

X	A	A	B	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 8

Y	D*	A	A	Y	X
Y	C	B	B	Y	Y
Y	C	A	B	X	X
X	A	A	B	X	X
X	D	E	A	Y	X
Y	A	E	A	Y	Y
Y	A	E	B	Y	Y
Y	C	D	E	X	X
X	D	D	B	X	X
X	D	D	B		

* I sent signals that I will play Y and in the long run we will all earn 99.

Subject 13

X	B	D	A	Y	X
X	A	C	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 17

Y	B	B	A	Y	Y
Y	C	B	B	X	X
X	A	B	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 23

X	A	A	C	Y	Y
X	B	A	C	X	Y
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 14

Y	B	A	A	X	X
Y	C	B	B	X	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	A	B	Y	Y
Y	C	B	B	X	X
X	A	B	B	X	X
X	A	B	B		

Subject 21

X	A	D	A	Y	Y
X	B	C	B	X	Y
Y	A	A	B	X	X
Y	A	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	E	B	Y	Y
Y	C	C	B	Y	Y
Y	A	C	B	Y	X
Y	A	A	B	X	X
X	A	A	B		

Subject 24

X	A	B	A	X	Y
X	A	B	A	X	X
X	B	A	C	Y	X
X	A	C	C	Y	Y
X	A	C	C	Y	Y
X	A	A	C	Y	X
X	A	A	A	Y	Y
X	A	C	A	X	Y
X	A	B	B	X	Y
X	A	B	B		

Subject 15

X	D*	B	B	X	Y
X	E*	A	C	Y	Y
X	E*	A	C	Y	Y
X	E*	C	C	Y	Y
X	B	E	C	X	Y
X	B	E	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C		

Subject 19

Y	B	B	C	X	Y
Y	C	B	C	Y	Y
Y	C	B	C	X	Y
Y	C	B	C	X	Y
Y	C	B	C	Y	Y
Y	C	B	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C		

Subject 22

Y	B	A	B	X	Y
Y	C	B	C	Y	Y
Y	C	B	C	X	Y
Y	C	B	C	X	Y
Y	C	B	C	Y	Y
Y	C	B	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	A	A	Y	Y
Y	C	A	A		

* I wanted to maximise my payout.

Subject 16

X	A	B	A	X	X
X	B	C	B	X	X
X	B	C	C	Y	Y
X	B	C	C	Y	Y
X	B	A	A	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 18

Y	B	A	A	Y	Y
Y	C	E	B	Y	Y
Y	C	E	A	Y	Y
Y	C	E	A	X	Y
X	B	E	A	X	X
X	B	E	A	X	X
X	B	A	A	X	X
X	B	A	A	X	X
X	B	A	A	X	X
X	B	A	A		

Subject 20

X	D	A	B	Y	X
X	B	B	C	X	Y
Y	A	B	C	X	Y
Y	A	B	C	X	Y
Y	A	B	A	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Phase 3

Subject 1

Y	B	A	B	X	Y
Y	C	B	C	Y	Y
Y	C	B	A	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X	A	B	A		

Subject 2

X	A	B	B	X	X
X	A	C	B	Y	Y
X	A	C	A	X	X
X	B	B	B	X	Y
X	B	B	B	X	Y
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 3

Y	B	B	A	Y	Y
Y	C	C	B	Y	X
X	A	C	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B		

Subject 4

Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C		

Subject 5

Y	B	B	B	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	X	X
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	B	A		

Subject 12

Y	B	B	B	Y	Y
Y	E*	C	C	Y	Y
Y	E*	C	C	Y	Y
Y	E*	C	C	Y	Y
Y	E*	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	A	C		

* Ensure Y trend continues.

Subject 6

X	A	B	B	X	X
X	B	C	C	X	X
X	B	A	C	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	C	X	X
X	A	C	C	X	X
X	A	C	A	X	X
X	A	A	A		

Subject 9

Y	C	D	D	Y	Y
Y	C	A	E	X	X
X	A	A	C	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	C	X	Y
Y	A	A	C	X	Y
Y	A	A	C	X	X
X	A	A	A		

Subject 7

Y	B	A	B	X	X
Y	C	B	C	Y	Y
Y	C	B	D	X	X
X	D	B	B	X	Y
X	D	B	B	X	X
X	D	B	A	X	X
Y	C	E	A	X	X
Y	C	B	C	Y	Y
Y	C	B	C	X	X
X	E*	B	A		

* Because the other two in this group do not understand cooperation, number two is just stupid.

Subject 8

Y	B	A	B	Y	Y
Y	C	B	C	X	X
Y	C	B	E	Y	Y
Y	C	B	C	Y	Y
Y	C	B	B	X	X
Y	C	B	B	Y	Y
Y	E*	B	C	Y	Y
Y	C	B	C	X	X
X	D	B	C	X	Y
X	B	B	C		

* So that we can all go to *Y* and earn 120.

Subject 10

Y	B	B	A	Y	X
Y	C	C	B	Y	Y
Y	C	A	B	Y	Y
Y	A	A	B	X	X
X	B	A	A	X	X
X	B	B	B	X	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	A	B	X	X
X	B	C	B		

Subject 11

X	A	B	B	Y	Y
X	A	B	A	Y	Y
X	A	A	A	X	Y
X	B	A	A	Y	Y
X	A	A	A	Y	X
X	A	A	A	X	X
X	A	C	C	Y	Y
X	A	A	A	Y	Y
X	A	A	A	X	X
X	A	C	A		

Subject 13

Y	B	A	A	Y	X
Y	C	B	B	X	X
Y	C	B	C	X	Y
Y	C	B	C	Y	Y
Y	A	B	A	X	X
X	A	B	A	X	Y
X	A	B	C	X	Y
Y	C	B	C	X	Y
Y	C	B	C	Y	Y
Y	C	B	C		

Subject 14

X	B	C	B	X	X
X	A	C	B	X	X
X	A	C	C	Y	Y
X	A	C	C	Y	Y
X	A	C	A	Y	X
X	A	A	A	X	X
X	A	A	C	Y	Y
X	A	C	C	Y	X
X	A	C	C	Y	Y
X	A	C	C		

Subject 15

X	D	B	B	Y	Y
X	B	C	A	X	X
Y	A	A	A	Y	Y
Y	A	A	E	Y	X
X	E*	A	E	X	X
X	E*	A	E	X	X
Y	A	E	E	X	Y
Y	A	C	B	Y	X
Y	A	C	B	Y	X
Y	A	C	B		

* Tried to maximise my payoff.

Subject 16

Y	B	B	A	Y	Y
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	X	X
X	B	B	B		

Subject 17

Y	B	B	C	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	X
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	X	X
X	A	A	A		

Subject 24

X	A	B	B	X	Y
X	B	C	C	Y	Y
X	B	C	C	Y	Y
X	B	E	C	Y	Y
X	B	E	E	Y	Y
X	B	A	A	Y	Y
X	A	A	A	Y	Y
X	A	E	E	Y	Y
X	A	A	A	Y	Y
X	A	B	B		

Subject 18

Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C		

Subject 21

Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C		

Subject 19

Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C		

Subject 20

Y	B	B	B	X	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	Y	Y
Y	A	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	X
X	E	B	C		

Subject 22

Y	B	B	B	Y	Y
Y	C	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	C	A	A	Y	Y
Y	C	A	C	X	X
X	B	B	C		

Subject 23

Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	Y
Y	C	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	Y
Y	C	B	B		

C.2 Second subsession's data

Phase 4

Subject 1

Y	C	C	C	Y	Y
Y	C	A	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	A	B	A		

Subject 4

Y	C	C	C	Y	Y
Y	C	C	A	X	Y
Y	C	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	B	B	B		

Subject 9

Y	C	C	C	Y	Y
Y	A	C	C	Y	X
Y	A	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	B	B	B		

Subject 2

X	B	C	C	Y	Y
X	C	C	C	X	Y
X	B	B	C	X	X
X	B	B	C	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 5

Y	A	B	C	Y	Y
Y	A	B	C	X	Y
X	A	B	C	X	X
X	A	B	C	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X	A	B	A	X	X
X		B	A		

Subject 10

Y	C	B	C	X	X
Y	C	B	C	X	X
Y	C	B	B	X	Y
Y	C	A	A	X	X
X	A	A	A	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	A	A	A		

Subject 7

X	D	A	C	X	X
X	B	A	C	X	X
X	B	A	C	X	Y
X	B	C	A	X	X
X	B	A	A	X	X
X	D	A	C	X	X
X	B	A	A	X	X
X	B	A	A	X	X
X	B	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 8

X	D	D	C	X	Y
X	D	D	C	X	X
X	D	D	C	X	Y
Y	C	D	A	X	X
X	D	D	C	X	X
X	B	D	C	X	X
X	B	D	B	X	X
X	D	B	B	X	X
X	D	B	B	X	X
X	D	B	B		

Subject 12

Y	E*	E	E	X	X
Y	E*	E	E	Y	X
Y	A	E	E	X	X
X	A	E	C	Y	X
Y	A	E	A	X	X
Y	C	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

* All Y greater than all X.

Subject 3

X	A	B	B	X	X
Y	E*	E	B	Y	Y
Y	C	C	B	Y	Y
Y	C	B	B	X	X
Y	C	B	B	Y	Y
Y	C	B	B	X	X
X	A	A	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	A	B		

* To change the behaviour of the others.

Subject 6

X	A	A	A	X	X
X	A	C	A	X	X
X	A	C	A	X	X
X	A	C	A	X	X
X	A	C	A	X	X
X	A	C	A	X	X
X	A	A	C	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B		

Subject 11

X	B	B	B	X	X
X	A	C	A	X	X
X	A	C	A	X	X
X	A	A	A	X	X
X	A	E	A	X	X
X	A	E	A	Y	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 13

Y	C	A	C	X	Y
Y	C	A	C	X	Y
Y	C	A	C	X	Y
Y	C	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B		

Subject 16

X	B	C	C	X	X
X	B	C	C	Y	Y
X	B	C	C	Y	Y
X	B	C	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 21

Y	C	C	B	Y	Y
Y	C	C	B	Y	X
Y	C	C	B	Y	X
X	A	C	B	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 14

Y	C	C	C	Y	Y
X	B	C	C	Y	Y
X	B	C	C	X	Y
X	A	C	A	X	Y
X	A	C	A	Y	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B		

Subject 17

Y	C	C	C	Y	Y
Y	C	B	A	Y	Y
Y	C	B	A	Y	Y
Y	C	E	A	X	Y
Y	C	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 22

Y	A	C	C	Y	Y
Y	A	B	C	Y	Y
Y	A	B	A	Y	Y
Y	A	A	A	X	Y
X	A	A	A	Y	Y
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B		

Subject 15

X	E*	E	E	X	X
X	E*	A	E	X	X
X	E*	A	E	X	X
X	A	A	E	X	X
X	A	A	E	X	X
X	A	A	E	X	X
X	A	A	E	X	X
X	A	A	E	X	X
X	A	A	E	X	X
X	A	A	E		

* I expected everyone else to chose X.

Subject 18

X	B	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B		

Subject 23

X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 19

Y	C	B	B	X	Y
Y	C	B	B	X	X
Y	C	B	B	X	X
X	A	B	B	X	X
Y	C	B	B	Y	X
Y	C	B	B	X	X
Y	C	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 20

X	D	C	B	Y	Y
X	E	C	B	Y	X
X	A	C	B	Y	X
Y	E	B	B	X	X
X	E	C	B	X	X
X	A	C	B	X	X
X	E	B	B	X	X
X	E	B	B	X	X
X	E	B	B	X	X
X	B	B	B		

Subject 24

X	B	A	B	X	X
X	B	A	B	Y	X
X	B	A	B	Y	X
X	B	A	A	X	X
X	B	A	A	X	Y
X	B	A	B	Y	X
X	B	D	B	Y	X
X	B	D	B	X	X
X	B	B	B	X	X
X	B	B	B		

Phase 5

Subject 1

Y	C	B	B	X	X
Y	A	B	B	Y	X
Y	C	B	B	X	X
X	A	A	B	X	X
Y	C	A	B	Y	X
Y	C	A	B	X	X
Y	C	A	B	X	X
Y	C	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B	X	X
X	A	A	B		

Subject 11

X	B	C	E	Y	Y
X	A	C	A	X	X
X	A	C	A	Y	X
Y	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 5

X	A	C	A	X	X
X	A	C	A	X	X
X	A	E	A	X	X
X	A	A	A	X	X
X	A	C	A	Y	Y
X	A	C	A	X	X
X	A	C	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 2

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	A	A	Y	Y
Y	C	A	A	X	X
Y	C	C	B		

Subject 9

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	B	C	C		

Subject 12

Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	E*	E	E		

* All Y greater than all X.

Subject 3

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	B	C	Y	Y
Y	C	B	C	X	Y
X	A	A	A	X	X
X	A	A	A		

Subject 7

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
X	B	C	C	Y	Y
X	B	C	C	X	X
X	B	A	C	X	X
X	A	A	A		

Subject 10

Y	B	B	B	Y	Y
Y	B	B	B	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	X
Y	A	A	B	X	X
X	B	B	B		

Subject 4

Y	C	B	C	X	X
Y	C	B	C	X	Y
Y	C	B	C	Y	Y
Y	C	B	D	X	X
X	D	B	D	X	X
X	B	B	B	X	X
X	D	B	C	X	X
X	D	B	C	X	X
X	D	B	C	X	X
X	D	B	C		

Subject 6

X	A	C	C	X	X
X	A	C	C	X	X
X	A	C	C	X	X
X	A	C	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	C	X	X
X	A	B	C	X	X
X	A	B	B	X	X
X	A	B	C		

Subject 8

Y	C	C	D	Y	X
Y	C	C	D	Y	X
Y	C	A	D	X	X
X	B	C	B	X	X
X	D	B	A	X	X
X	A	A	A	X	X
Y	C	A	A	X	X
Y	C	A	A	Y	X
Y	A	A	A	Y	Y
Y	C	A	A		

Subject 13

Y	C	C	C	Y	X
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	A	A	Y	X
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	X
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	B		

Subject 17

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	A	A	X	X
Y	C	C	C	Y	Y
Y	C	C	A	Y	X
Y	C	C	E	Y	Y
Y	C	C	E	X	X
X	A	C	E		

Subject 23

Y	C	C	C	Y	X
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	Y
X	B	C	A		

Subject 14

X	B	B	B	Y	X
X	B	A	B	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	Y
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A		

Subject 21

X	D	B	B	Y	Y
X	D	C	B	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	Y	Y
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	E	E	X	X
X	A	E	E	X	X
X	A	E	E		

Subject 24

X	B	B	D	X	X
X	B	C	D	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	D	X	X
X	D	B	D	X	X
X	D	B	B	X	X
X	D	B	B	X	X
X	B	B	B		

Subject 15

Y	A	E	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	X	X
X	A	A	B		

Subject 19

Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	X
Y	C	A	B		

Subject 22

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	B	B	C		

Subject 16

X	B	C	B	X	X
X	B	C	B	Y	X
X	B	C	B	Y	X
X	B	C	B	Y	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B		

Subject 18

Y	C	B	B	Y	X
Y	C	B	B	X	Y
Y	C	B	B	Y	X
Y	C	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B	X	X
X	A	B	B		

Subject 20

X	E	B	C	X	Y
X	E	B	C	X	X
X	E	B	C	X	X
X	E	B	C	X	Y
X	D	B	A	X	X
X	D	B	A	X	X
X	D	B	B	X	X
X	D	B	B	X	X
X	D	B	B	X	X
X	D	B	B		

Phase 6

Subject 1

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	A	C	X	X
Y	A	A	A	X	X
X	A	A	A		

Subject 2

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	X	X
X	B	B	B		

Subject 3

Y	C	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	A	A	E	Y	X
Y	C	A	E	X	X
X	A	A	E		

Subject 4

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	E	X	X
X	B	B	B		

Subject 5

Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	B	X	X
X	A	A	A		

Subject 12

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	A	A	A		

Subject 6

X	B	B	B	X	X
X	B	B	B	X	X
X	B	B	B	X	X
X	A	A	A	X	X
X	B	B	B	X	X
X	B	A	B	X	X
X	A	B	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	D	A	A		

Subject 9

X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	B	A	A		

Subject 7

X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	A	A	A	X	X
X	D	A	A	X	X
X	D	A	A		

Subject 8

Y	C	D	D	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	B	Y	Y
Y	C	B	B	Y	Y
Y	C	B	B	X	X
X	A	B	C		

Subject 10

X	B	B	A	X	X
Y	A	A	A	Y	X
Y	C	A	A	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	A	C	Y	Y
Y	C	B	B		

Subject 11

Y	C	C	B	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
X	B	A	A		

Subject 13

Y	C	C	C	X	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	B	X	X
X	A	B	B		

Subject 14

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	A	C	C	Y	X
Y	A	C	B	Y	X
X	B	A	B		

Subject 15

Y	E*	E	E	Y	Y
Y	E*	E	E	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	E	Y	X
X	A	A	E	X	X
X	A	E	E		

* I expected everyone else to choose Y.

Subject 16

Y	C	B	C	Y	Y
Y	C	B	C	X	Y
Y	C	B	C	X	X
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	B		

Subject 17

X	B	C	C	Y	Y
X	A	C	C	Y	Y
X	A	C	A	X	Y
Y	C	C	A	Y	X
Y	C	C	A	Y	X
Y	C	C	A	Y	X
Y	C	C	A	Y	X
Y	C	C	A	Y	X
Y	C	C	A		

Subject 24

Y	A	A	B	Y	Y
Y	A	A	B	Y	Y
Y	C	A	B	Y	Y
Y	C	A	A	Y	Y
Y	C	A	A	Y	Y
Y	C	A	A	Y	Y
Y	C	A	A	Y	Y
Y	C	A	A	Y	Y
X	B	C	C		

Subject 18

Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	Y	Y
Y	C	C	C	X	X
X	A	C	C		

Subject 21

Y	E	E	E	Y	Y
Y	A	E	E	Y	Y
Y	A	E	E	Y	Y
Y	A	E	E	Y	Y
Y	A	E	E	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	A	Y	Y
Y	A	A	C	X	X
X	A	A	A		

Subject 19

Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	Y	Y
Y	A	A	C	X	X
X	A	B	B		

Subject 20

Y	C	C	C	Y	Y
Y	E	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	X
X	D	B	A	X	Y
Y	E	B	A	X	Y
Y	A	C	A	X	X
X	D	B	A	X	X
X	E	B	B		

Subject 22

Y	C	C	C	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	Y	Y
Y	C	C	B	X	Y
X	A	B	C	X	X
X	A	C	C	Y	X
X	A	C	B	X	X
X	A	B	B	X	X
X	A	B	B		

Subject 23

Y	A	A	A	Y	Y
X	B	C	C	X	X
X	B	C	C	X	Y
X	B	C	C	X	Y
X	B	C	C	Y	Y
Y	A	A	A	Y	X
Y	C	A	A	X	X
X	B	A	C	X	X
X	B	B	B	X	X
X	B	B	B		

Bibliography

Althammer, W. & Buchholz, W. (1993), 'Lindhal–Equilibria as the Outcome of a Non–Cooperative Game', *European Journal of Political Economy* **9**, 399–405.

Andreoni, J. (1988), 'Why Free Ride? Strategies and Learning in Public Good Experiments', *Journal of Public Economics* **37**, 291–304.

Andreoni, J. (1989), 'Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence', *Journal of Public Economics* **97**(6), 1447–1458.

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

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

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

Andreoni, J. (1995*b*), 'Warm–Glow versus Cold–Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments', *The Quarterly Journal of Economics* **110**, 1–21.

Andreoni, J. & Bergstrom, T. (1996), 'Do Government Subsidies Increase the Private Supply of Public Goods?', *Public Choice* **88**, 295–308.

Andreoni, J. & Miller, J. H. (1993), 'Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence', *Economic Journal* **103**, 570–585.

Aumann, R. & Hart, S. (1992), *Handbook of Game Theory*, Amsterdam: North–Holland.

Aumann, R. J. (1976), 'Agreeing to Disagree', *Annals of Statistics* **4**, 1236–1239.

Bacharach, M. (1985), 'Some Extensions to a Claim of Aumann in an Axiomatic Model of Knowledge.', *Journal of Economic Theory* **37**, 167–190.

Bagnoli, M., Ben-David, S. & McKee, M. (1992), 'Voluntary Provision of Public Goods: The Multiple Unit Case.', *Journal of Public Economics* **47**, 85–106.

Bagnoli, M. & Lipman, B. L. (1989), 'Provision of Public Goods: Fully Implementing the Core Through Private Contributions', *Review of Economic Studies* **56**, 583–601.

Bagnoli, M. & McKee, M. (1991), 'Voluntary Contribution Games: Efficient Private Provision of Public Goods', *Economic Inquiry* **29**, 351–365.

Becker, G. (1974), 'A Theory of Social Interactions', *Journal of Political Economy* **82**, 1063–1093.

Becker, G. (1976), *The Economic Approach to Human Behaviour*, Chicago: Chicago University Press.

Bergstrom, T. (1995), 'On the Evolution of Altruistic Ethical Rules for Siblings', *American Economic Review* **85**, 58–81.

Bergstrom, T., Blume, L. & Varian, H. (1986), 'On the Private Provision of Public Goods', *Journal of Public Economics* **29**, 25–49.

Bergstrom, T. & Stark, O. (1993), 'How Altruism Can Prevail in an Evolutionary Environment', *American Economic Review* **83**, 149–155.

Bernheim, B. D. (1984), 'Rationalizable Strategic Behavior', *Econometrica* **52**, 1007–1028.

Bernheim, B. D. (1986), 'On the Voluntary and Involuntary Provision of Public Goods', *American Economic Review* **76**, 789–793.

Bernoulli, D. (1738), 'Exposition of a New Theory of the Measurement of Risk'. English translation in *Econometrica*, 22(1954):23–36.

Bester, H. & Güth, W. (1998), 'Is Altruism Evolutionary Stable?', *Journal of Economic Behavior and Organization* **34**, 193–209.

Binmore, K. G. (1990), *Essays on the Foundations of Game Theory*, Oxford: Basil Blackwell Ltd.

Binmore, K. G. (1993), *Playing Fair: Game Theory and the Social Contract*, Cambridge, MA: MIT Press.

Blount, S. (1995), 'When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences', *Organizational Behavior and Human Decision Processes* **63**, 131–144.

Bohm, P. (1972), 'Estimating Demand for Public Goods: An Experiment', *European Economic Review* **3**, 111–130.

Bolton, G. E. (1991), 'A Comparative Model of Bargaining: Theory and Evidence', *American Economic Review* **81**(5), 1096–1136.

Bolton, G. E. (1998), 'Bargaining and Dilemma Games: From Laboratory Data Towards Theoretical Synthesis', *Experimental Economics* **1**, 257–281.

Bolton, G. E., Brandts, J. & Katok, E. (1997), A Simple Test of Explanations for Contributions in Dilemma Games. Discussion paper 360.96, Institut d'Anàlisi Econòmica (CSIC), Barcelona.

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

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

Borel, E. (1921), 'La Théorie Du Jeu et Les Équations Intégrales à Noyau Symétrique', *Comptes Rendus de l'Académie des Sciences* **173**, 1304–1308.

Boylan, R. T. (1990), Equilibria Resistant to Mutation. Social Science Working Paper 691, California Institute of Technology.

Brandts, J. & Schram, A. (1997), Cooperation and Noise in Public Goods Experiments: Applying the Contribution Function Approach. Working Paper, Institut d'Anàlisi Econòmica (CSIC), Barcelona.

Brewer, M. B. & Kramer, R. M. (1986), 'Choice Behavior in Social Dilemmas: Effects of Social Identity, Group Size, and Decision Framing', *Journal of Personality and Social Psychology* **50**, 543–545.

Brunner, J. K. & Falkinger, J. (1995), Nonneutrality of Taxes and Subsidies for the Private Provision of Public Goods. Arbeitspapier No. 9519, Department of Economics, Linz.

Buchanan, J. M. (1971), *The Bases for Collective Action*, New York: General Learning Press.

Burlando, R. & Hey, J. D. (1997), 'Do Anglosaxons Free-Ride More?', *Journal of Public Economics* **64**, 41–60.

Camerer, C. (1995), Individual Decision Making, *in* J. H. Kagel & A. E. Roth, eds, 'The Handbook of Experimental Economics', Princeton, NJ: Princeton University Press.

Camerer, C., Knez, M. & Weber, R. (1996), Virtual Observability. Working Paper, California Institute of Technology.

Carter, J. R. & Irons, M. D. (1991), 'Are Economists Different, and If so, Why?', *Journal of Economic Perspectives* **5**, 171–177.

Chakrabarti, S., Lord, W. & Rangazas, P. (1993), 'Uncertain Altruism and Investment in Children', *American Economic Review* **83**, 994–1002.

Chan, K. R., Godby, S., Mestelman, S. & Muller, A. (1994), Boundary Effects and Voluntary Contributions to Public Goods. Working Paper, Department of Economics, McMaster University.

Charness, G. & Haruvy, E. (1999), Altruism, Equity and Reciprocity in a Gift–Exchange Experiment: An Encompassing Approach. Working Paper, Universitat Pompeu Fabra and University of Texas.

Chen, Y. & Plott, C. R. (1996), 'The Groves–Ledyard Mechanism: An Experimental Study of Institutional Design', *Journal of Public Economics* **59**, 335–364.

Coate, S. (1995), 'Altruism, the Samaritan's Dilemma, and Government Transfer Policy', *American Economic Review* **85**, 46–57.

Collard, D. (1978), *Altruism and Economy*, Oxford: Martin Robertson.

Collard, D. (1983), 'Economics of Philanthropy: A Comment', *Economic Journal* **93**, 637–638.

Crawford, V. P. & Haller, H. (1990), 'Learning How to Cooperate: Optimal Play in Repeated Coordination Games', *Econometrica* **58**(3), 571–575.

Croson, R. (1996), 'Partners and Strangers Revisited', *Economics Letters* **53**, 25–32.

Croson, R. (1998*a*), 'Effects of Eliciting Beliefs in a Linear Public Goods Game', OPIM Working Paper, The Wharton School of the University of Pennsylvania, Philadelphia.

Croson, R. (1998*b*), 'Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games', OPIM Working Paper, The Wharton School of the University of Pennsylvania, Philadelphia.

Cross, J. (1980), Some Comments on the Papers by Kagel and Battalio and by Smith, *in* J. Kmenta & J. Ramsey, eds, 'Evaluation of Econometric Models', New York: Academic Press.

Danzinger, L. & Schnytzer, A. (1991), 'Implementing the Lindhal Voluntary–Exchange Mechanism', *European Journal of Political Economy* **7**, 55–64.

Davis, D. D. & Holt, C. A. (1993), *Experimental Economics*, Princeton, NJ: Princeton University Press.

Dawes, R. M. (1975), Formal Models of Dilemmas in Social Decision–Making, *in* M. F. Kaplan & S. Schwartz, eds, 'Human Judgement and Decision Processes: Formal and Mathematical Approaches', New York: Academic Press.

Dawes, R. M. (1980), 'Social Dilemmas', *Annual Review of Psychology* **31**, 169–193.

Dawes, R. M., McTavish, J. & Shaklee, H. (1977), 'Behavior, Communication, and Assumptions About Other People's Behavior in a Commons Dilemma Situation', *Journal of Personality and Social Psychology* **35**(1), 1–11.

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

Downs, A. (1957), *An Economic Theory of Democracy*, New York: Harper & Row.

Duffy, J. & Feltovich, N. (forthcoming), 'Does Observation of Others Affect Learning in Strategic Environments? An Experimental Study', *International Journal of Game Theory* .

Edgeworth, F. Y. (1881), *Mathematical Physics*, London: P. Kegan.

Elster, J. (1982), 'Marxism, Functionalism and Game Theory', *Theory and Society* **11**, 453–482.

Falkinger, J. (1994), 'The Private Provision of Public Goods When the Relative Size of Contributions Matters', *Finanzarchiv* **51**, 358–371.

Falkinger, J. (1996), 'Efficient Private Provision of Public Goods by Rewarding Deviations from Average', *Journal of Public Economics* **62**, 413–422.

Falkinger, J., Fehr, E., Gächter, S. & Winter-Ebmer, R. (1999), A Simple Mechanism for the Efficient Provision of Public Goods–Experimental Evidence. Working Paper No. 3, Institute for Empirical Research in Economics, University of Zurich.

Fehr, E., Kirchsteiger, G. & Riedl, A. (1993), 'Does Fairness Prevent Market Clearing? An Experimental Investigation', *Quarterly Journal of Economics* **108**, 437–459.

Fehr, E. & Schmidt, K. M. (1999), 'A Theory of Fairness, Competition, and Cooperation', *Quarterly Journal of Economics* **114**, 817–868.

Festinger, L. (1957), *A Theory of Cognitive Dissonance*, Stanford: Stanford University Press.

Fiorina, M. P. & Plott, C. R. (1978), 'Committee Decisions under Majority Rule: An Experimental Study', *American Political Science Review* **72**, 575–598.

Fleishman, J. (1988), 'The Effects of Decision Framing and Others' Behavior on Cooperation in a Social Dilemma', *Journal of Conflict Resolution* **32**, 162–180.

Friedman, D. & Sunder, S. (1994), *Experimental Methods: A Primer for Economists*, New York: Cambridge University Press.

Friedman, J. W. (1986), *Game Theory with Applications to Economics*, New York: Oxford University Press.

Friedman, M. (1953), *Essays in Positive Economics*, Chicago: University of Chicago Press.

Friedman, M. & Savage, L. J. (1948), 'The Utility Analysis of Choices Involving Risk', *Journal of Political Economy* **56**, 279–304.

Gale, J., Binmore, K. G. & Samuelson, L. (1995), 'Learning to Be Imperfect: The Ultimatum Game', *Games and Economic Behavior* **8**, 56–90.

Geanakoplos, J., Pearce, D. & Stacchetti, E. (1989), 'Psychological Games and Sequential Rationality', *Games and Economic Behavior* **1**, 60–79.

Gomes, M. C., Crawford, V. P. & Broseta, B. (1996), Experimental Studies of Strategic Sophistication and Cognition in Normal–Form Games. Working Paper, UCSD and Arizona University.

Grether, D. M. (1992), 'Testing Bayes Rule and the Representativeness Heuristic: Some Experimental Results', *Journal of Economic Behavior and Organization* **17**, 31–57.

Groves, T. & Ledyard, J. (1977), 'Optimal Allocation of Public Goods: A Solution to the "Free Rider" Problem', *Econometrica* **45**, 783–809.

Guttman, J. M. (1978), 'Understanding Collective Action: Matching Behavior', *American Economic Review* **68**, 251–255.

Guttman, J. M. (1986), 'Matching Behavior and Collective Action: Some Experimental Evidence', *Journal of Economic Behavior and Organization* **7**, 171–198.

Guttman, J. M. (1987), 'A Non–Cournot Model of Voluntary Collective Action', *Economica* **54**, 1–19.

Hardin, G. R. (1968), 'The Tragedy of the Commons', *Science* **162**, 1243–1248.

Hardin, G. R. (1976), 'Carrying Capacity as an Ethical Concept', *Soundings: Interdiscip. Journal* **59**, 121–137.

Hargreaves-Heap, S. P. & Varoufakis, Y. (1995), *Game Theory: A Critical Introduction*, London: Routledge.

Harrison, G. (1994), 'Expected Utility and the Experimentalists', *Empirical Economics* **19**, 223–253.

Harrison, G. & Hirshleifer, J. (1989), 'An Experimental Evaluation of Weakest Link/Best Shot Models of Public Goods', *Journal of Political Economy* **97**, 201–225.

Harsanyi, J. C. (1967-68), 'Games with Incomplete Information Played by 'Bayesian' Players', *Management Science* **14**, 159–182, 320–334, 486–502.

Harsanyi, J. C. (1980), 'Rule Utilitarianism, Rights, Obligations and the Theory of Rational Behaviour', *Theory and Decision* **12**, 115–133.

Hey, J. D. (1991), *Experiments in Economics*, Oxford: Basil Blackwell Ltd.

Hory, H. (1992), 'Utility Functionals with Nonpaternalistic Intergenerational Altruism: The Case Where Altruism Extends to Many Generations', *Journal of Economic Theory* **56**, 451–467.

Isaac, R. M., McCue, K. & Plott, C. R. (1985), 'Public Goods Provision in an Experimental Environment', *Journal of Public Economics* **26**, 51–74.

Isaac, R. M. & Walker, J. (1988), 'Communication and Free Riding Behavior: The Voluntary Contribution Mechanism', *Economic Inquiry* **26**, 585–608.

Isaac, R. M. & Walker, J. (1992), Nash as an Organizing Principle in the Voluntary Contribution of Public Goods. Working Paper in Economics, Indiana University.

Isaac, R. M., Walker, J. & Thomas, S. (1984), 'Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations', *Public Choice* **43**(1), 113–149.

Kalai, E. & Lehrer, E. (1990), Rational Learning Leads to Nash Equilibrium. Discussion Paper 895, Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston, Ill.

Kelley, H. & Stahelski, A. (1970), 'Social Interaction Basis of Cooperators' and Competitors' Beliefs About Others', *Journal of Personality and Social Psychology* **16**, 66–91.

Keser, C. (1996), 'Voluntary Contributions to a Public Good When Partial Contribution is a Dominant Strategy', *Economics Letters* **50**, 359–366.

Keser, C. (1997), SUPER: Strategies Used in Public Goods Experimentation Rounds. Working Paper 97/24, Sonderforschungsbereich 504, University of Mannheim.

Keser, C. (1999), Strategically Planned Behavior in Public Goods Experiments: Should We Rely More on Voluntary Contributions to Public Goods? Working Paper, University of Karlsruhe and CIRANO, Montréal.

Keser, C. & van Winden, F. (2000), 'Conditional Cooperation and Voluntary Contributions to Public Goods', *Scandinavian Journal of Economics* **102**, 23–39.

Kim, O. & Walker, M. (1984), 'The Free Rider Problem: Experimental Evidence', *Public Choice* **43**, 3–24.

Koopmans, T. C. (1957), *Three Essays on the State of Economic Science*, New York: McGraw–Hill.

Kreps, D. M. (1990), *Game Theory and Economic Modelling*, Oxford: Clarendon Press.

Kreps, D. M., Milgrom, P., Roberts, J. & Wilson, R. (1982), 'Rational Cooperation in the Finitely Repeated Prisoners' Dilemma', *Journal of Economic Theory* **27**, 245–252.

Kreps, D. M. & Wilson, R. (1982), 'Reputation and Imperfect Information', *Journal of Economic Theory* **27**, 253–279.

Kuhlman, D. M. & Wimberley, D. (1976), 'Expectations of Choice Behavior Held by Cooperators, Competitors and Individualists Across Four Classes of Experimental Games', *Journal of Personality and Social Psychology* **3**, 69–81.

Laffont, J. J. (1975), 'Macroeconomics Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics', *Economica* **42**, 430–437.

Laffont, J. J. (1987), Incentives and the Allocation of Public Goods, *in* A. J. Auerbach & M. Feldstein, eds, 'Handbook of Public Economics', Amsterdam: North–Holland.

Ledyard, J. O. (1995), Public Goods: A Survey of Experimental Research, *in* J. H. Kagel & A. E. Roth, eds, 'The Handbook of Experimental Economics', Princeton, NJ: Princeton University Press.

Levine, D. K. (1998), 'Modeling Altruism and Spitefulness in Experiments', *Review of Economic Dynamics* **1**, 593–622.

Lewis, D. (1969), *Convention*, Cambridge, MA: Harvard University Press.

Lipsey, R. & Crystal, K. (1995), *Positive Economics*, New York: Oxford University Press.

Loomes, G. (1991), Experimental Methods in Economics, *in* D. Greenaway, M. Bleaney & I. Stewart, eds, 'Companion to Contemporary Economic Thought', London: Routledge.

Marwell, G. & Ames, R. (1979), 'Experiments on the Provision of Public Goods I: Resources, Interest, Group Size, and the Free–Rider Problem', *American Journal of Sociology* **84**(6), 1335–1360.

Marwell, G. & Ames, R. (1980), 'Experiments on the Provision of Public Goods II: Provision Points, Stakes, Experience, and the Free–Rider Problem', *American Journal of Sociology* **85**(4), 926–937.

Marwell, G. & Ames, R. (1981), 'Economists Free Ride, Does Anyone Else? Experiments on the Provision of Public Goods', *Journal of Public Economics* **15**, 295–310.

McKelvey, R. D. & Palfrey, T. R. (1992), 'An Experimental Study of the Centipede Game', *Econometrica* **60**, 803–836.

Messé, L. & Sivacek, J. (1979), 'Predictions of Others' Responses in a Mixed–Motive Game: Self–Justification or False Consensus?', *Journal of Personality and Social Psychology* **37**, 602–607.

Messick, D., Wilke, H., Brewer, M., Kramer, R., Zemke, P. E. & Lui, L. (1983), 'Individual Adaptations and Structural Change as Solutions to Social Dilemmas', *Journal of Personality and Social Psychology* **44**, 294–309.

Milgrom, P. & Roberts, J. (1982), 'Predation, Reputation and Entry Deterrence', *Journal of Economic Theory* **27**, 280–312.

Miller, J. H. & Andreoni, J. (1991), 'Can Evolutionary Dynamics Explain Free–Riding in Experiments?', *Economics Letters* **36**, 9–15.

Myerson, R. B. (1997), *Game Theory: Analysis of Conflict*, Cambridge, MA: Harvard University Press.

Nash, J. F. (1951), 'Noncooperative Games', *Annals of Mathematics* **54**, 289–295.

Offerman, T. (1996), Beliefs and Decision Rules in Public Good Games, PhD thesis, University of Amsterdam. Also available as Tinbergen Institute Research Series no. 124.

Offerman, T. & Schram, A. (1993), Selfishness, Rationality and Orientation Re–Examined: Evidence from Economic and Psychological Experiments. Working Paper, University of Amsterdam.

Offerman, T. & Sonnemans, J. (1995), Learning by Experience and Learning by Imitating Successful Others. Working Paper, University of Amsterdam.

Offerman, T., Sonnemans, J. & Schram, A. (1996), 'Value Orientations, Expectations and Voluntary Contributions in Public Goods', *Economic Journal* **106**, 817–845.

Palfrey, T. R. & Prisbrey, J. E. (1992), Anomalous Behavior in Linear Public Goods Experiments: How Much and Why? Social Science Working Paper 833, California Institute of Technology.

Palfrey, T. R. & Prisbrey, J. E. (1996), 'Altruism, Reputation and Noise in Linear Public Goods Experiments', *Journal of Public Economics* **61**, 409–427.

Palfrey, T. R. & Prisbrey, J. E. (1997), 'Anomalous Behavior in Public Goods Experiments: How Much and Why?', *American Economic Review* **87**, 829–846.

Palfrey, T. R. & Rosenthal, H. (1991), Testing Game–Theoretical Models of Free Riding: New Evidence on Probability Bias and Learning, *in* T. Palfrey, ed., 'Laboratory Research in Political Economy', Ann Arbor: University of Michigan Press.

Pearce, D. G. (1984), 'Rationalizable Strategic Behavior and the Problem of Perfection', *Econometrica* **52**, 1029–1050.

Plott, C. R. (1982), 'Industrial Organization Theory and Experimental Economics', *Journal of Economic Literature* **20**, 1485–1527.

Plott, C. R. (1987), Dimensions of Parallelism: Some Policy Applications of Experimental Methods, *in* A. E. Roth, ed., 'Laboratory Experimentation in Economics: Six Points of View', New York: Cambridge University Press.

Plott, C. R. (1991), 'Will Economics Become an Experimental Science?', *Southern Economic Journal* **57**, 901–919.

Poppe, M. & Utens, L. (1986), 'Effects of Greed and Fear of Being Gypped in a Social Dilemma Situation with Changing Pool Size', *Journal of Economic Psychology* **7**, 61–73.

Prasniskar, V. & Roth, A. E. (1992), 'Consideration of Fairness and Strategy: Experimental Data from Sequential Games', *Quarterly Journal of Economics* **107**, 865–888.

Rabin, M. (1993), 'Incorporating Fairness Into Game Theory and Economics', *American Economic Review* **83**(5), 1281–1302.

Ramsey, F. P. (1926), 'Truth and Probability'. Reprinted in Kyburg, Jr. H. E. and Smokler, H., editors, *Studies in Subjective Probability* (1964), pp. 62–92. New York: Wiley.

Rangazas, P. (1991), 'Human Capital Investment in Wealth–Constrained Families with Two–Sided Altruism', *Economics Letters* **35**, 137–141.

Roth, A. (1987), *Laboratory Experimentation in Economics: Six Points of View*, New York: Cambridge University Press.

Roth, A. E. & Erev, I. (1995), 'Learning in Extensive–Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term', *Games and Economic Behavior* **8**, 164–212.

Samuelson, J. D. & Nordhaus, W. D. (1985), *Principles of Economics*, New York: McGraw-Hill (12th ed.).

Samuelson, L. (1989), Dominated Strategies and Common Knowledge. Pennsylvania State University Working Paper.

Samuelson, P. A. (1954), 'The Pure Theory of Public Expenditure', *Review of Economics and Statistics* **36**, 387–389.

Samuelson, P. A. (1993), 'Altruism as a Problem Involving Group versus Individual Selection in Economics and Biology', *American Economic Review* **83**, 143–148.

Savage, L. J. (1954), *The Foundations of Statistics*, New York: Wiley.

Schroeder, D., Jensen, T., Reed, A., Sullivan, D. & Schwab, M. (1983), 'The Actions of Others as Determinants of Behavior in Social Trap Situations', *Journal of Experimental Social Psychology* **19**, 522–539.

Sefton, M. & Steinberg, R. (1996), 'Reward Structures in Public Good Experiments', *Journal of Public Economics* **61**, 263–287.

Sell, J. & Wilson, R. K. (1991), 'Levels of Information and Contributions to Public Goods', *Social Forces* **70**(1), 107–124.

Selten, R. (1965), 'Spieltheorische Behandlung Eines Oligopolmodells mit Nachfragetragheit', *Zeitschrift fuer die gesampte Staatswissenschaft* **121**, 301–324, 667–689.

Selten, R. (1975), 'Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory* **4**, 25–55.

Selten, R. (1978), 'The Chain-Store Paradox', *Theory and Decision* **9**, 127–159.

Selten, R. (1994), Descriptive Approaches to Cooperation. Discussion Paper B–292, University of Bonn.

Shafir, E. (1994), 'Uncertainty and the Difficulty of Thinking Through Disjunctions', *Cognition* **50**, 403–430.

Shafir, E. & Tversky, A. (1992), 'Thinking Through Uncertainty: Non Consequential Reasoning and Choice', *Cognitive Psychology* **24**, 449–474.

Simon, H. A. (1993), 'Altruism and Economics', *American Economic Review* **83**, 156–161.

Smith, V., Kehoe, M. & Cremer, M. (1995), 'The Private Provision of Public Goods: Altruism and Voluntary Giving', *Journal of Public Economics* **58**, 107–126.

Smith, V. L. (1976), 'Experimental Economics: Induced Value Theory', *American Economic Review* **66**(2), 274–279.

Smith, V. L. (1980), Relevance of Laboratory Experiments to Testing Resource Allocation Theory, *in* J. Kmenta & J. Ramsey, eds, 'Evaluation of Econometric Models', New York: Academic Press.

Smith, V. L. (1982), 'Microeconomic Systems as an Experimental Science', *American Economic Review* **72**, 923–955.

Smith, V. L. (1989), 'Theory, Experiments and Economics', *Journal of Economic Perspectives* **3**, 151–169.

Smith, V. L. & Walker, J. M. (1993), 'Monetary Rewards and Decision Costs in Experimental Economics', *Economic Inquiry* **31**, 245–261.

Starmer, C. (1996), Experiments in Economics ... (Should We Trust the Dismal Scientists in White Coats?). Working Paper, School of Economic and Social Studies, UEA, Norwich.

Strawcyznski, M. (1994), 'Government Intervention as a Bequest Substitute', *Journal of Public Economics* **53**, 477–495.

Sugden, R. (1982), 'On the Economics of Philanthropy', *Economic Journal* **92**, 341–350.

Sugden, R. (1984), 'Reciprocity: The Supply of Public Goods Through Voluntary Contributions', *Economic Journal* **94**, 772–787.

Thurstone, L. (1931), 'The Indifference Function', *Journal of Social Psychology* **2**, 139–167.

Unger, L. (1991), 'Altruism as a Motivation to Volunteer', *Journal of Economic Psychology* **12**, 71–100.

Varian, H. R. (1994*a*), 'A Solution to the Problem of Externalities When Agents are Well–Informed', *American Economic Review* **84**, 1278–1293.

Varian, H. R. (1994*b*), 'Sequential Contributions to Public Goods', *Journal of Public Economics* **53**, 165–186.

von Neumann, J. (1928), 'Zur Theorie der Gesellschaftsspiele', *Mathematische Annalen* **100**, 295–320. English translation in Luce, R. D. and Tucker, A. W., editors, *Contributions to the Theory of Games* IV (1959), pp. 13–42. Princeton, NJ: Princeton University Press.

von Neumann, J. & Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press. Second Ed., 1947.

Wallis, W. & Friedman, M. (1942), The Empirical Derivation of Indifference Functions, *in* O. Lange, F. McIntyre & T. Yntema, eds, 'Studies in Mathematical Economics and Econometrics in Memory of Henry Schults', Chicago: University of Chicago Press.

Warr, P. G. (1982), 'Pareto Optimal Redistribution and Private Charity', *Journal of Public Economics* **19**, 131–138.

Warr, P. G. (1983), 'The Private Provision of a Public Good is Independent of the Distribution of Income', *Economic Letters* **13**, 207–211.

Weimann, J. (1994), 'Individual Behaviour in a Free Riding Experiment', *Journal of Public Economics* **54**, 185–200.

Wilde, L. (1980), On the Use of Laboratory Experiments in Economics, *in* J. Pitt, ed., 'The Philosophy of Economics', Dordrecht: Reidel.

Wong, S. (1987), Positive Economics, *in* J. Eatwell, M. Milgate & P. Newman, eds, 'The New Palgrave Dictionary of Economics', London: The Macmillan Press Limited.

Zermelo, E. (1913), 'Uber eine Anwendung der Mengenlehre Auf Die Theorie Des Schachspiels', *Proceeding Fifth International Congress of Mathematicians* **2**, 501–504.

THESIS CONTAINS CD/DVD