

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

Foundations for Cooperation in Social Dilemmas

Toh-Kyeong Ahn

**Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in Political Science
Indiana University**

August 2001

UMI Number: 3024232

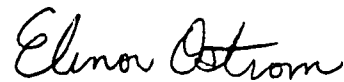
UMI[®]

UMI Microform 3024232

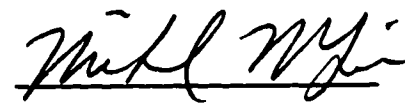
Copyright 2001 by Bell & Howell Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

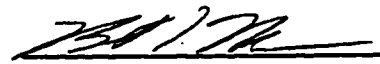


Elinor Ostrom, Ph. D., Chair

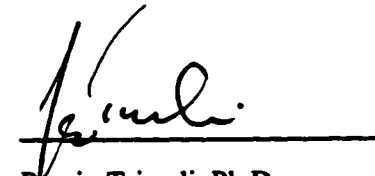


Mike McGinnis, Ph. D.

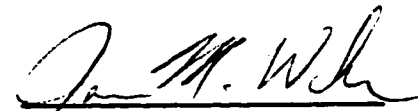
Doctoral Committee



Burt Monroe, Ph. D.



Pravin Trivedi, Ph.D.



James Walker, Ph.D.

November 30, 2000

Toh-Kyeong Ahn

Foundations for Cooperation in Social Dilemmas

Institutional approaches to social dilemmas have so far focused on how to create incentive structures that channel individuals' behavior into socially desirable outcomes assuming that everyone is selfish. However, the presumption of universal selfishness is not only empirically invalid but may also result in inefficient policy prescriptions in the short run and a culture of distrust in the long run. The game theoretic models that depart from the universal self-interest assumption, on the other hand, have not been successful in incorporating both individual rationality and inter-individual heterogeneity.

This dissertation develops game theoretic models of social dilemmas with individuals who are rational in the sense that they have preferences and try to maximize expected utility but heterogeneous in the sense that they have different preferences over possible social outcomes. This dissertation tests the implications of the models using a series of experimental data sets. The empirical results are: (1) there is a significant proportion of individuals who are rational but not selfish, (2) non-selfish individuals' motivations are best described on a dimension of equity/fairness rather than being thought of as unconditional altruism, and (3) the possibility of sustained mutual cooperation in finitely repeated social dilemmas is affected by the material conditions of the action situation, the composition of types within the population, and individuals' willingness to take risks to initiate coordination.

Acknowledgements

Professors Elinor Ostrom and James M. Walker taught and supported me far beyond the scope of this dissertation. From them, I have learned how to teach a course, how to run an experiment, how to write an article, and how to work together with others. They set standards for me that I will strive to meet for the rest of my academic life. I am deeply indebted to Professors Elinor and Vincent Ostrom for the care and concern they have exercised over the course of my studies, for my professional and personal well-being.

I owe my expertise and enthusiasm for statistics to Professor Pravin K. Trivedi who made me realize how shallow my understanding of statistical methodology was and how entertaining and illuminating its application in the Social Sciences can be. Professors Burt Monroe, David Schmidt, and Eric Rasmusen taught me game theory, thus, provided me with a lifetime asset. Professor Monroe has always impressed me with his superb ability to explain the essentials in accessible ways to non-specialists. Professor Schmidt has been my friendly colleague, as well as teacher. He also helped me greatly to improve the game-theoretic models in this dissertation. Professor Mike McGinnis provided me with many valuable comments during the course of my working on this dissertation.

Myeun Kim, Sangbae Kim, Nathan Basik, Mike Craw, and Ray Eliason have provided me with their friendship and much practical assistance during my graduate years.

The colleagues and staff at the Workshop in Political Theory and Policy Analysis, Indiana University have provided me with a friendly and stimulating environment for the past several years.

Patty Zielinski's editorial support improved the readability of this dissertation greatly.

Contents

List of Figures	viii
List of Tables	x
1 Introduction: Solutions to Social Dilemmas and Individual Motivations	1
1.1 Institutional Solutions to Social Dilemmas	5
1.1.1 State or Entrepreneurial Solutions	7
1.1.2 Market/Privatization Solution	11
1.1.3 Self-Governance	12
1.2 Non-Institutional Solutions and Endogeneity of Preference	16
1.2.1 Non-Institutional Solutions to Social Dilemmas and the Existence of Multiple Types of Individuals	16
1.2.2 Endogeneity of Preferences and the Role of Institutions in Cultural Evolution	20
1.3 The Plan of the Dissertation	23
2 Definition of Social Dilemmas and the Selfishness Assumption	28
2.1 Introduction	28
2.2 Defining Social Dilemmas	30
2.2.1 Social Dilemmas and Collective Action Problems	30
2.2.2 The Common Definition of Social Dilemma	31
2.2.3 Definition of Social Dilemmas in Terms of Utilities	33
2.2.4 The Nature of von Neumann-Morgenstern Utility	35
2.2.5 Standard Non-cooperative Game Theoretical Approach to Social Dilemmas	41
2.2.6 A Behavioral Definition of Social Dilemma	45
2.3 A Critical Review of the Selfishness Assumption	49
2.4 The Need for a Behavioral Framework	55
3 Frameworks and Models	59
3.1 Introduction	59
3.2 Review of Frameworks: MES and IAD	62

3.2.1	V. Smith's Microeconomic System Framework	62
3.2.2	Institutional Analysis and Development Framework	69
3.3	A Behavioral Framework for Social Dilemmas and Its Operations	71
3.3.1	A Behavioral Framework of Social Dilemmas	72
3.3.2	Operation of the Working Parts in the Framework	74
3.3.3	Theories and Inference on Motivation	78
3.4	Generic Utility Function	82
3.4.1	Motivational Factors: A Generic Utility Function	82
3.4.2	Arguments in Utility Function: Factors Affecting Preference	84
3.5	Non-selfish Utility Functions	86
3.5.1	Altruism	87
3.5.2	Inequity Aversion	87
3.5.3	ERC (Equity, Reciprocity, Competition)	88
3.5.4	Fairness	89
3.5.5	Relationship Accounting	91
4	Altruism or Equity?	92
4.1	Introduction	92
4.2	Preference-Ordering Types	99
4.2.1	Altruism	101
4.2.2	Inequity Aversion	107
4.3	Equilibria	111
4.3.1	Inequity Aversion Model	113
4.3.2	Altruism Model	126
4.4	Behavior in Four Information Sets	135
4.4.1	Inequity Aversion Model	136
4.4.2	Altruism Model	138
4.5	Empirical Tests and Results	140
4.5.1	Data	140
4.5.2	Types of Preference Orderings	143
4.5.3	Behavior in Four Information Sets	147
4.6	Conclusion	150
5	Finite Repetition of a 2×2 Social Dilemma	152
5.1	Introduction	152
5.2	Finitely Repeated 2×2 Social Dilemma Game with Uncertainty	154
5.3	Cooperative Equilibria of the Finitely Repeated Game with Two Types of Players	156
5.3.1	Cooperative Equilibrium of the Finitely Repeated Game with a Gen- eral Distribution of Types	163
5.3.2	Hybrid Equilibria of the Finitely Repeated Game	169
5.4	Conclusion	170

6 Heterogeneity and Interdependence	172
6.1 Introduction	172
6.2 Experimental Design and Procedure	176
6.3 Overall Results	179
6.4 Heterogeneity: Fixed Effects Logit Analyses	190
6.5 Interdependence of Strategies: Bivariate Probit Analyses	203
6.6 Conclusion	209
7 Conclusion: Heterogeneous Motivations and Cooperation in Social Dilemmas	212

List of Figures

1.1	Social Dilemma Action Situation	5
1.2	Social Dilemma with External Sanction	7
1.3	Social Dilemma with External Sanction: Probabilistic Enforcement	9
1.4	Privatization Solution	12
1.5	Social Dilemma with Self-Governance	13
1.6	2×2 Social Dilemma <i>Game</i> with Guilt	18
2.1	The Same Game?	40
2.2	Stage Game Matrix of the Finitely Repeated Prisoner's Dilemma Game. Source: Kreps et al.(1982:245).	44
3.1	Dual representation of an Environment	63
3.2	Components of Action Arenas. Source: Ostrom, Gardner, and Walker (1994: 29).	70
4.1	Indifference Mapping of Pure Selfishness	98
4.2	Indifference Mapping of Linear Altruism	98
4.3	Indifference Mapping of Inequity Aversion	98
4.4	2×2 Social Dilemma	99
4.5	Normal Form Game Representation of a 2×2 Social Dilemma	101
4.6	Monetary Incentive Structure of AOW Experiment	141
4.7	Decision Problem in SURVEY	143
5.1	2×2 Social Dilemma Action Situation	154
5.2	Von Neumann-Morgenstern Utilities in the 2×2 Social Dilemma Game	155
6.1	Material Payoff Structure of Stage Games (<i>Source: Schmidt et al., forthcoming</i>)	178
6.2	Frequency of <i>Cooperation</i> across Stages	180
6.3	Decisions Made by a Pair of Players	181
6.4	Decisions in Phase 1, Session 1	182
6.5	Decisions in Phase 1, Session 2	182
6.6	Decisions in Phase 1, Session 3	183
6.7	Decisions in Phase 2 Low Cooperators' Group, Session 1	183

6.8	Decisions in Phase 2 Low Cooperators' Group, Session 2	184
6.9	Decisions in Phase 2 Low Cooperators' Group, Session 3	184
6.10	Decisions in Phase 2 High Cooperators' Group, Session 1	185
6.11	Decisions in Phase 2 High Cooperators' Group, Session 2	185
6.12	Decisions in Phase 2 High Cooperators' Group, Session 3	186

List of Tables

2.1	Cardinalization of preference over two events: xPy	36
2.2	Cardinalization of preference over three events: $xPyPz$	36
4.1	Preference Types and Interpretation: Altruism Model	107
4.2	Conditions for Preference Types: Altruism Model	107
4.3	Preference Types and Interpretation: Inequity Aversion Model	111
4.4	Conditions for Preference Types: Inequity Aversion Model	111
4.5	Preference Types: Altruism Model with $\pi (F_n > G_n)$	127
4.6	Equilibrium of a 2×2 Social Dilemma Game : Altruism Model with $\pi :$ ($F_n < G_n$)	131
4.7	Preference-Ordering Possibilities: Inequity Aversion Model	145
4.8	Preference-Ordering Possibilities: Altruism Model	145
4.9	Distribution of Types: Inequity Aversion Model	146
4.10	Distribution of Types: Altruism Model	146
4.11	Marginal Explanatory Power of Three Models	147
4.12	Frequency of <i>Cooperation</i> in Four Information Sets	148
4.13	Frequency of <i>Cooperation</i> in Four Information Sets: A Comparison of Studies	150
6.1	Normalized Material Payoff Parameters	179
6.2	Variables and Explanation	195
6.3	Cooperation: Logit and Conditional Fixed Effects Logit Estimates	196
6.4	The Frequency of Cooperation across Game Structures	197
6.5	Phase 2 High Group Individuals' Behavior in Phases 1 and 2	200
6.6	Cooperation in Phases 1 and 2: By Phase 2 Group	201
6.7	Correlation between Paired Players' Choices	204
6.8	Cooperation: Pooled Bivariate Probit Estimates	207
6.9	Cooperation: Pooled Bivariate Probit Estimates with Individual Dummies .	210

Chapter 1

Introduction: Solutions to Social Dilemmas and Individual Motivations

Many social opportunities are squandered when individuals act in a narrow short-term selfish manner. Opportunities are also foregone when all citizens are assumed to act in a narrow selfish manner by policymakers who attempt to prevent the squandering. Herders in the famous example by Hardin (1968) are presumed to overgraze their jointly used pasture leading to “the tragedy of the commons.” An external authority who assumes that the herders are all selfish, and thus will inevitably overgraze the pasture when left alone, may also prevent the herders from devising their own rules to utilize the pasture more efficiently. Design of an institution to cope with the possible tragedy of the commons, be it a set of self-governing rules designed by the herders themselves or imposed rules

by an external authority, needs to be based on an empirically valid understanding of the motivations and behavioral rules held by the herders. Institutions, thus designed, reshape behavior of the herders in the short term and their norms and motivations in the long run.

This dissertation investigates the motivational foundation of cooperation in social dilemmas. A social dilemma is an action situation involving a group of individuals and commonly valued good(s) in which if every individual tries to maximize their material reward, an outcome results in which all the individuals are worse off than they could be if they were to adopt at least one alternative strategy available to each of them. Do individuals in social dilemma situations inevitably and persistently act selfishly as they are supposed to do in markets? If selfishness is not the proper approximation of all individuals' motivations, what are the empirically valid alternatives? Can the alternative models cope with the fact that individuals differ from each other, a fact that has hindered incorporation of non-selfish motivations into game theoretic models for so long? Are the alternative motivational models of non-selfishness testable empirically? Can they pass the tests?

Both normative and empirical concerns require a serious re-examination of the selfishness assumption in the study of social dilemmas. The very survival and existence of human civilization suggest that human beings, throughout history, have been quite successful in dealing with the pervasive presence of social dilemmas in everyday life. Participation in civil rights movements and democratic institutions, including voting, also make us suspicious of the presumption that all individuals pursue mainly their own self-interest. Though not every individual participates in those movements or institutions, the level of participation is high enough to make a significant political impact and to sustain basic institutions

of democracy.

Controlled laboratory experiments on social dilemmas also cast a heavy doubt on the validity of the universal selfishness assumption that characterizes all individuals in all situations as pursuers of only one's own material well-being. Experiments of various game forms including the prisoner's dilemma (Hayashi et al., 1999; Ahn et al., 2001; Clark and Sefton, 2001), common-pool resources (Walker, Gardner, and Ostrom, 1990), public goods provision (Dawes, McTavish, and Shaklee, 1977; Marwell and Ames, 1979; Isaac, Walker, and Thomas, 1984), ultimatum (Camerer and Thaler, 1995; Hoffman, McCabe, and Smith, 1996; Cameron, 1995), dictator (Davis and Holt, 1993: 263-9; Forsythe, Horowitz, and Savin, 1994), and gift-exchange (Fehr, Kirschsteiger, and Riedls, 1993; Berg, Dickhaut, and McCabe, 1995) games, have shown that a significant proportion of subjects do consciously choose a course of action that does not maximize their expected material payoffs from the game (for a succinct review, see Ostrom, 2000).

The findings are puzzling for two reasons. First, it appears as though the underlying motivations that drive human behavior differ in two broad sets of experimental action situations: markets and social dilemmas. In the market experiments, predictions based on a self-interest assumption have been quite successful (Smith, 1991). Efforts to develop a theory that explains this apparent inconsistency have been launched in recent years, but the width and depth of the problem have not yet allowed any single theory to resolve all, or even at least the key, intermingled issues (Rabin, 1993; Crawford and Ostrom, 1995; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Falk and Fischbacher, 1998; Dufwenberg and Kirchsteiger, 1998). Second, the extent of non-selfish actions seems to be systematically

affected by factors that should not play any role when individuals act selfishly (Ostrom, 1998; Schmidt et al., forthcoming). Schmidt et al. categorize these factors into four broad sets: structure of pecuniary benefits, player types, information about player types, and the linkage among players in repeated games. The systematic influence of these institutional and environmental factors on behavior in social dilemmas suggests that there is a rational decision-making mechanism held by individuals, other than a blind pursuit of self-interest.

Normative concerns relate to the endogeneity of motivations. When individuals are assumed to be selfish, *everyone* is assumed as such and everyone is assumed as such *all the time*. The possibility of the change of preferences, evolution of motivations, and the role of institutions in the evolutionary dynamics is precluded. Motivational factors become constants in institutional design and, thus, the goal of institutional design is limited to inducing desired static outcomes given the static motivations. However, when an institution is designed based on a set of presumptions regarding individual motivation, it tends to prove itself in the long run by remolding the original motivation of the people into one presumed by the designers of the institutions. An institution based on the presumption that individuals are invariably opportunistic may transform originally well-natured individuals into a group of opportunists.

The other side of the story is that wisely devised institutions cannot only redirect immediate behavior of individuals to socially beneficial ones, but also reshape social norms and culture through an evolutionary process. Aristotle's (1962) remarks on the purpose of legislation summarizes this dimension of institutions succinctly: "Lawgivers make the citizen good by inculcating habits in them, and this is the aim of every lawgiver; if he does

not succeed in doing that, his legislation is a failure. It is in this that a good constitution differs from a bad one" (quoted in Bowles, 1999: 1).

1.1 Institutional Solutions to Social Dilemmas

Figure 1.1 presents an action situation involving two individuals. The action situation involves a structure of material payoffs that parallel the utility payoff structure of the Prisoner's Dilemma game. Since the payoffs are material, the matrix of Figure 1.1 does not constitute a game. Normal form representation of the action situation requires assumptions regarding how individuals transform the material payoffs into utilities.

If the action situation is not repeated, and the individuals try to maximize their material payoffs, the action situation can be modeled as the Prisoner's Dilemma in which both individuals choose *Defection*. The resulting outcome is (P, P) , which is less preferred to (R, R) by both of them. The dilemma of the action situation is that an outcome unambiguously less preferred by the members of a group may result from the very fact that each individual in the group tries to achieve an outcome more preferred by him. On the other hand, one can also argue that the individuals are given a chance to achieve (R, R) . Thus they have an opportunity to achieve a mutually beneficial social outcome.

		<i>Player 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Player 1</i>	<i>Cooperation</i>	<i>R, R</i>	<i>S, T</i>
	<i>Defection</i>	<i>P, P</i>	<i>T, S</i>

**T, R, P, and S* represent material payoffs
 ** $T > R > P > S$

Figure 1.1: Social Dilemma Action Situation

The action situation extracts the essence of the problem involved in various types of what have been called so far social dilemmas (Dawes, 1975; 1980), collective action problems (Olson, 1965), or the tragedy of the commons (Hardin, 1968). To start thinking about the institutional solutions to this problem, it is necessary to put the abstract situation into a substantive context. The grazing problem faced by the herders involves a lot more complicated factors. For example, actions available for the herders are rarely a dichotomy of *Cooperation* or *Defection*. But the nature of the problem – the conflict between individual and collective interest – remains the same. One could understand *Cooperation* as the kind of action by the herders (choice of the number of cattle a herder chooses to feed on the pasture) that maximizes joint return (measured, for example, by the number of successfully raised cattle or the market value of them). *Defection* then corresponds to a choice of action that maximizes an individual herder's return given that the other herder also tries to maximize the same value. If the two herders both try to maximize their individual material returns from the pasture, then each will choose the individually rational number of cattle on the pasture and the outcome will be an overgrazing of the pasture.

This situation is ubiquitous whenever an individual's rational action imposes a larger externality to others than the benefit he receives from it. Many cases of natural resource appropriation, decisions that affect environmental quality, and behavior in the pursuit of commonly valued social or political causes have at their core the problem described above.

While market interactions are generally believed to produce socially desirable consequences based on individually selfish choices, that is not the case with social dilemmas.

Prosperity and sometimes the very existence and survival of a society depends, in large part on how individuals deal with the problem collectively, or how individuals prevent the possible squandering and realize the productive opportunity. Many social opportunities involve actively creating action situations of which the essence is the social dilemma. A confident group of individuals would find and create the opportunity while a pessimistic one may not even notice the chance.

1.1.1 State or Entrepreneurial Solutions

Figure 1.2 represents a modified social dilemma action situation in which an external authority imposes punishment in the form of a fine to the defectors. An external government imposes a fine of c to defectors to make *Cooperation* a rational choice for a selfish individual. The magnitude of the fine c , then, has to be big enough such that $T - c < R$.

Governmental interventions to various social action situations are based on this rather simple idea. This idea also extends to the most basic principle of the organization of social order. In Hobbes's (1960) *Leviathan*, individuals transfer all their natural rights to the State to avoid the seemingly inevitable outcome of (P, P) in Figure 1.1 and to achieve (R, R) in Figure 1.2.

		<i>Player 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Player 1</i>	<i>Cooperation</i>	R, R	$S, T - c$
	<i>Defection</i>	$T - c, S$	$P - c, P - c$

* $T - c < R$

Figure 1.2: Social Dilemma with External Sanction

Alchian and Demsetz (1972) put a similar idea at the core of rationalizing the necessity of economic enterprises owned by a residual claimant. They argue that individuals do not have an incentive to do their best to maximize group production in a teamwork because individual input is not exactly measurable and thus the co-workers share the joint profit not based on individual input. Alchian and Demsetz argue that the predominant form of modern production, private ownership and the separation of ownership and work, is the solution to this dilemma of teamwork. The manager (owner) monitors individual input, which is a measure of *Cooperation/Defection*, and rewards or punishes individual members in the work team accordingly. In essence, therefore, Alchian and Demsetz's rationalization is on the exact same track as Hobbes's rationalization of the absolute state.¹

Insofar as the external authority performs its monitoring and sanctioning roles efficiently and fairly, this solution would not constitute a problem. Governmental activities at various levels on various social problems often generate this beneficial impact. However, in principle as well as in reality, the solution by external authority is not satisfactory. Two intermingled problems stand out. First, there is the problem of fair implementation. Unfair implementation of the power by the external authority can occur in different ways. The external authority can collude with one of the two players and participate in the exploitation of the other. Or, when the power granted to the external authority is unchecked, it can use its power to exploit both of the players. The problems do not vanish all together even when the external authority is determined to function in a fair and public-minded manner. First of all, the establishment and maintenance of an external authority requires a cost bearing

¹Ostrom and Hennessey (1975: 23) argue that "Alchian and Demsetz's common contracting party is, in effect, the sovereign in a sole proprietorship." Bowles (1985) refers to Alchian and Demsetz's model as "neo-Hobbesian."

		<i>Player 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Player 1</i>	<i>Cooperation</i>	$R - \beta c, R - \beta c$	$S - \beta c, T - \alpha c$
	<i>Defection</i>	$T - \alpha c, S - \beta c$	$P - \alpha c, P - \alpha c$

$$0 < \beta < \alpha < 1$$

α : the probability of punishing Defection

β : the probability of punishing Cooperation

c : fine imposed to those who are perceived to be Defectors

Figure 1.3: Social Dilemma with External Sanction: Probabilistic Enforcement

on the part of the players. Second, there usually are local exigencies that cannot be easily captured in a simplified presentation of action situations (see Figure 1.1). What constitutes *Cooperation* and *Defection* are not always clear to external observers, thus creating the possibility of honest, but quite costly, mistakes on the part of the external authority.

Figure 1.3 represents a social dilemma with external enforcement when the external authority's enforcement of the rule is probabilistic (see also Ostrom, 1990: 8-12). Probabilistic enforcement can result from many reasons including the impure motivation of the authority, lack of information, or any kind of mistake involved in enforcement of the rules. In the modified action situation, the external authority can make mistakes, as can the authorities in real world. It sanctions *Defection* only with a probability α . Further, it punishes *Cooperation* with a positive probability β .² It is assumed that two probabilities are independent and correct punishment is more likely than incorrect punishment ($\alpha > \beta$).

Analysis of the game in Figure 1.3 shows a case in which the intervention by the external authority may not guarantee social cooperation. Ensuring cooperation requires that, when the other player cooperates, the payoff from a player's *Cooperation* is higher

²Figure 1.3 is not intended as *the* model of governmental intervention in social dilemmas. It is just one, but quite typical, model. For example, one might wish to include the cost of maintaining the external authority. In that case, additional reduction of all payoff entries is necessary.

than that from *Defection*. Or,

$$R - \beta c > T - \alpha c \quad (1.1)$$

$$c > \frac{T - R}{\alpha - \beta}. \quad (1.2)$$

Only when the relationship among the size of punishment (c), the enforcement probabilities (α, β), and the material payoff entries (T, R) expressed in (1.2) holds, *Cooperation* becomes the dominant choice for a selfish individual. When this relationship does not hold, individuals may still find themselves trapped in (*Defection, Defection*) even after they give up their rights to act freely under the situation.

Or, in another possible scenario, *Cooperation* is ensured by severe punishment, but due to the combined effect of a relatively high probability of punishing cooperators and a large amount of imposed fine, individuals may end up worse off in the state-guaranteed (*Cooperation, Cooperation*) than they could do in (*Defection, Defection*) when there is no external authority. Inequality (1.4) specifies this condition.

$$P > R - \beta c \quad (1.3)$$

$$\beta > \frac{R - P}{c}. \quad (1.4)$$

1.1.2 Market/Privatization Solution

For many social dilemmas, privatization is a possible solution. For example, in the case of Hardin's two herders, privatization would mean, among other alternatives, that the two individuals divide the land evenly and let each feed his cattle only on his own land. If the soils are similar and rainfall is regular, this could lead to a better utilization of the pasture since each individual herder has to absorb any consequences of his or her own overgrazing. The problem related to privatization solution is case specific. In the case of the herder's dilemma, possible problems are (1) cost of privatization: fences cost money to build and maintain, (2) the lost chance of utilizing scale economies, and (3) the inability to share risk if rainfall is patchy and unpredictable.

In general, the essence of the problem seems to be the under-utilization of scale economies and the capacity to share risk. In the case of a teamwork dilemma, one way of privatization is not to form a team at all and let each worker work for himself. In that case, the workers would not only eliminate the potential dilemma involved in teamwork, but also the productive opportunity provided by it. When privatization is a plausible option and it involves foregone opportunities to utilize scale economies, the outcome of privatization is often a certain material payoff of which the magnitude is smaller than R (lost opportunity) but greater than P (avoided dilemma). In sum, privatization is a recommendable solution only when it is physically possible and there is no hope for achieving voluntary cooperation.

Figure 1.4 shows a modified social dilemma action situation via privatization. Action situations resulting from privatization vary depending on the nature of the original social dilemma and the way privatization is done. Therefore, except for the fact that

		<i>Player 2</i>		
		<i>Cooperation</i>	<i>Privatization</i>	<i>Defection</i>
<i>Player 1</i>	<i>Cooperation</i>	<i>R, R</i>	<i>Q, Q</i>	<i>S, T</i>
	<i>Privatization</i>	<i>Q, Q</i>	<i>Q, Q</i>	<i>Q, Q</i>
	<i>Defection</i>	<i>T, S</i>	<i>Q, Q</i>	<i>P, P</i>

**(T > R > Q > P > S)*

Figure 1.4: Privatization Solution

what can be secured by privatization is less than when individuals successfully cooperate within the social dilemma, the exact presentation of the modified action situation would vary. Though Figure 1.4 shows a timeless action situation, it can be better understood if a decision regarding privatization or not is assumed to be made in the first stage. If and only if both individuals agree not to privatize, then the social dilemma game in Figure 1.1 will be played. If at least one of the two players prefers privatization, the subsequent material payoff for either of the players is Q , which is smaller than R but bigger than P .

The beneficial aspect of privatization is that the individuals can escape the trap of mutual defection (P, P). On the other hand, the cost is the lost chance of utilizing the social opportunity of achieving a mutually beneficial outcome of (R, R). If at least one of the two individuals believes that the other will not cooperate, and if they fail to arrange higher-level institutional devise to facilitate self-governance, it seems inevitable that individuals would opt for privatization when it is technically and legally available.

1.1.3 Self-Governance

In natural settings, researchers often find that individuals involved in social dilemmas successfully deal with the problem resorting neither to external authorities nor to a private property market mechanism. Based on extensive empirical studies, Ostrom (1990)

		<i>Player 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Player 1</i>	<i>Cooperation</i>	$R - e/2, R - e/2$	$S - e/2, T - e/2 - c'$
	<i>Defection</i>	$T - e/2 - c', S - e/2$	$P - e/2, P - e/2$

* e : cost of establishing and maintaining self-governing institution
** c' : sanction to a defector

Figure 1.5: Social Dilemma with Self-Governance

has identified a set of design principles that facilitate self-governance. The essence of the self-governing solution is that the actors in the social dilemmas collectively devise an institutional arrangement with which to channel individual choices to collectively beneficial outcomes.

Compared to the market solution, the key advantage of self-governance is that it keeps the productive opportunity open. In terms of pure incentive schemes, self-governance often resembles the state solution in that it involves monitoring and sanctioning that makes *Cooperation* consistent with long-term self-interest. However, the self-governance solution has the advantage over the state solution in that the individuals, as a collective entity, retain the ultimate decision-making authority. In addition, when the individuals involved in a social dilemma devise and implement the rules of monitoring by themselves, they can utilize local knowledge better than external authorities.

Figure 1.5 presents a modified social dilemma action situation with a self-governing institution in the form of sanctioning to defectors.³ In the representation of the state solution

³See Ostrom (1990:15-18) for extensive form non-cooperative game representation of the self-governing solution. As with the case for the preceding figures, Figure 1.5 is not meant to be a normal form non-cooperative game. It simply is a matrix representation of an action situation, since the payoffs are not von Neuman-Morgenstern utilities but material payoffs. With the assumption that individuals' utilities are one-to-one mapping of their material payoffs, the normal form game representation of the action situation becomes identical to the matrix representation of the action situation. However, that is the assumption to be challenged in this study.

in Figure 1.3, cost of enforcement was not included assuming that the part of the fine levied on perceived defectors would be used as the cost of enforcement, which includes the costs of monitoring, sanctioning, and the maintenance of the external authority. Here, the costs are included explicitly. The costs include those of initial transactions to devise the rules, ongoing monitoring, and sanctioning. Design of rules, monitoring, and sanctioning may require involvement of an external actor. But the external actor differs from the external authority in the state solution since the external actor in the self-governance solution is only a part of the overall institutional arrangement devised by the individuals initially involved in the social dilemma. The external actor, when necessary, is an agent that can be hired and fired by the individuals.

In Figure 1.5, e is the cost of establishing and maintaining the self-governing institutional arrangement and c' is the sanction to a *Defector*. When both individuals defect, it is assumed that there will be no sanction, but this does not make an analytical difference. The condition to secure mutual *Cooperation* is that the material payoff for a player from *Cooperation* be greater than that from *Defection*, given that the other player also cooperates. In that case, the dilemma problem transforms into a coordination problem in which achieving mutual cooperation is not difficult in the presence of communication.

It also involves a certain formal requirement in the form of the relationship between the cost of self-governance and the magnitude of sanction:

$$R - e/2 > T - e/2 - c' \quad (1.5)$$

$$c' > T - R. \quad (1.6)$$

In other words, the expected sanction to a defector needs to be greater than the expected material gain from *Defection*, which is quite intuitive. A more important condition is not apparent in Figure 1.5 alone. The cost of self-governing should be reasonably small. Otherwise, what can be achieved by self-governing arrangements could be less than what individuals can secure when both defect in the absence of self-governance.

$$R - e/2 > P \quad (1.7)$$

$$e/2 < R - P. \quad (1.8)$$

In inequality (1.8), $R - P$ is the magnitude of the material gain each individual can achieve by means of devising self-governing arrangements. $e/2$ is an individual's share in the cost of establishing and maintaining the self-governing institution. Inequality (1.8) suggests, first of all, that the expected gains from mutual cooperation secured by self-governing institutions needs to be large enough to justify spending of one's share in the cost for the self-governance to succeed. With a moderate magnitude of expected gain from mutual cooperation, the cost of establishing and maintaining self-governing institutions need to be affordable for the individuals. A part of the cost is determined by the material/natural aspect of the dilemma without leaving much room for the individuals to choose. However, another, more social part of the cost exists that can vary substantially, providing a series of options from which the individuals can choose. In a sense, $e/2$ is the budget within the constraint of which individuals hoping to devise a self-governing arrangement need to maneuver.

After the first kind of cost is subtracted from the budget, whether or not, how

much, and in what manner individuals agree to spend the rest of the budget affects the success and failure and the gains from the self-governance. Individuals could agree to hire a monitor or file a legal contract in an external court system. When the budget is tightly limited, individuals cannot afford as much investment in monitoring and implementation activities. A tight budget needs careful allocation.

One of the key questions to be addressed to further develop the theory of self-governance is how to design efficient self-governing institutions with the limited resources that can be spent for establishing and maintaining those institutions. An institution with a high level of joint monitoring, accurate sanctioning, and complete information sharing would perform well. However, that kind of “perfect” institution is, more often than not, beyond the resources individuals can profitably spend. Here again, a correct understanding of motivations/behavioral rules of the individuals involved in a social dilemma becomes crucial in devising an efficient self-governing institution.

1.2 Non-Institutional Solutions and Endogeneity of Preference

1.2.1 Non-Institutional Solutions to Social Dilemmas and the Existence of Multiple Types of Individuals

The solutions examined thus far are all based on the assumption that individuals are fundamentally selfish. In the state solution, the external authority assumes that the two individuals are selfish and thus not capable of solving the dilemma themselves. In the

market/privatization solution, each of the two individuals believes that the other individual is selfish and thus there is no chance of achieving mutual cooperation. In a self-governing solution, individuals take the productive opportunity by devising their own rules, but still, at least in the example given in Figure 1.5, the basic presumption is that the individuals are selfish and thus only with a set of rules that guarantees that the material return from *Cooperation* is bigger than that from *Defection*, can they escape the dilemma.

The primary interest of this section is the ramification of the existence of multiple types of players, regularly observed in the experimental laboratory, to the design of self-governing institutions. Types of individuals in experimental settings can be identified based on the extent to which experimental subject's behavior and/or answers to questionnaires deviate from the predictions based on the universal selfishness assumption.

For example, in the dictator games in which each subject is asked to divide a normalized sum of 1 between him or herself and another player, a simple way of assigning the subject's type is θ ($0 \leq \theta \leq 1$), the amount he or she allocates to the other player. Or, the second movers in a sequential Prisoner's Dilemma experiment can be divided into two types of *reciprocator* and *egoist* depending on whether they cooperate in return when the first mover cooperates. Or, when a survey method is used that asks subjects to order the four possible outcomes of a *Prisoner's Dilemma* game according to their preferences (for example, Ahn, Ostrom, Walker, 2000), the ways the subjects order the outcomes are proxies of their types.

In all three examples, there should be only one type of player if everyone is selfish. The existence of a significant proportion of player types other than the one predicted by the

		<i>Player 1</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Player 2</i>	<i>Cooperation</i>	<i>R</i> , <i>R</i>	<i>S</i> , <i>T - g</i>
	<i>Defection</i>	<i>T - g</i> , <i>S</i>	<i>P - g</i> , <i>P - g</i>

Figure 1.6: 2×2 Social Dilemma Game with Guilt

standard assumption indicates that the universal selfishness assumption is not valid. The transformation of material payoffs into utilities by the individuals in social dilemmas and the methods of formalizing the transformation are the main topics of this dissertation and will be discussed in the following chapters.

Imagine that the two herders know each other very well and they know the size of cattle to optimally utilize the pasture. Also suppose that each of them regards it a great shame to raise cattle that are more than one's fair share, and each knows that the other herder feels the same. In other words, each of the two herders prefers to raise his fair share of the optimal number of cattle even though they know that adding some more cattle would increase their material return. In this case, the normal form game representation of the social dilemma differs from the one based only on material rewards. One way to represent the sense of guilt and the preference for fair behavior is by including, in the payoffs, a loss of utility associated with the choice of *Defection*.⁴

While Figures 1.1 to 1.5 were matrix representations of action situations in which the payoff entries were material rewards, Figure 1.6 is a normal form game constructed based on the basic social dilemma action situation (shown in Figure 1.1) in which payoff entries are von Neumann-Morgenstern utilities. Material payoffs of the game are the same

⁴Crawford and Ostrom's (1995) delta (δ) parameter models individuals' internalized norm in a similar way.

as those in the basic social dilemma. Feeling guilty makes an option less preferable, which is expressed in terms of utility by subtracting g from material payoff T . The condition that an individual's sense of guilt is strong enough to make *Defection* no longer the dominant strategy can be expressed as

$$R > T - g \quad (1.9)$$

$$g > T - R. \quad (1.10)$$

In other words, the sense of guilt associated with *Defection* (g) is bigger than the increased satisfaction of receiving higher material payoff ($T - R$) when one individual exploits the other person's *Cooperation*. The normal form game representation also assumes that players' preferences are common knowledge: that each individual knows the other's preference and he knows that the other individual knows that he knows, etc. In that case, the game is one of *Coordination*, not *Prisoner's Dilemma*. Especially with communication, mutual cooperation is not difficult for them. Many cases of successful cooperation in experimental social dilemmas, or successful collective actions observed in field settings where a small number of individuals cooperate on a face-to-face basis are likely to have been achieved this way. This most optimistic case can be called a non-institutional solution to social dilemmas.

However, one cannot rely solely on the goodwill of the people when thinking of the solutions to social dilemmas. In reality, it is more likely that only a proportion of individuals are willing to cooperate when others cooperate, while still a significant proportion

of individuals act as pure egoists. Most cases of self-governance will still require changes in institutional arrangements to achieve desired social goals. But the nature and focus of the self-governing institution would be quite different when there is a significant proportion of reciprocators than when most of the individuals are entirely egoistic.

In sum, limited resources for institutional building require that a self-governing institution be built on the empirically valid understanding of the motivations involved in a social dilemma. Whether all individuals are considered selfish, or the existence of multiple motivations is acknowledged, can make a big difference in the design of self-governing institutions.

1.2.2 Endogeneity of Preferences and the Role of Institutions in Cultural Evolution

In addition to the problem of cost, endogeneity of preferences is another concern that designers of institutions need to address. Endogeneity of preferences refers to the dynamics between institutions and individuals' norms and motivations. When viewed from the perspective of a group, the distribution of the kinds of norms held by individuals constitutes what can be called the culture of that group. For example, in the context of a dictator game in which a dictator i unilaterally decides how to divide a normalized sum of 1 between himself ($1 - \theta_i$) and another person (θ_i), θ_i is one way to denote individual i 's type and the cumulative distribution function $F(\theta_i)$ is the corresponding way of denoting the *culture* of the group. An institution can transform the culture of a population, and the latter again influences the kind of institutions the population needs in dealing with social dilemma problems.

For the purpose of simple presentation, let us assume that there are two types of individuals, conditional cooperators and opportunists, in a population involved in a social dilemma. Individuals in the population play variants of the basic game in Figure 1.1. The conditional cooperator type cooperates if the partner cooperates. Opportunists defect whenever doing so gives them a higher objective payoff. The evolutionary argument is that the proliferation of a type depends on the objective material payoff it receives on average from the game.

The dynamics between institution and culture result from the beneficial/detrimental impact of natural or artificial rules on the reproduction of each type. Among those institutional characteristics are: whether the game is played sequentially or simultaneously; whether information regarding one's partner is provided and if so on what level; whether the game is played only between two individuals or among the members of bigger subgroups of the population; whether the matching for a game is random or fixed; etc. Investigation of the impacts of these factors on the cultural evolution requires extensive theoretical and empirical research.

When the evolutionary implications of institutions are taken into account, institutional design needs to address long-term cultural consequences as well as the immediate behavioral impact of institutions (Ostrom, 2000). Several theoretical and empirical studies exist that examine the impacts of subsets of institutional characteristics on cultural evolution. Güth, Kliemt, and Peleg (2000) show that, when the game is played sequentially, the informational condition as well as initial distribution of types is critical in the proliferation of conditional cooperators. Theoretical analyses by Bohnet, Frey, and Huck (2001) show

that when the game is played sequentially with no information provided regarding partner's type the level of law enforcement has non-linear impact on the growth of conditionally cooperative individuals. When the logic of the survival of the fittest expands to intergroup competition, a group with institutions that are not quite beneficial to the growth of conditionally cooperative types could still survive inter-group competition if it initially has a significant proportion of cooperators (Bowles, 1999).

The dynamic relationship between the culture and institution of a group has multiple dimensions. In a static perspective of institutional design, the culture of a group at a given time affects the optimal design of self-governance. (Below, I stands for self-governing Institution.)

$$F_t(\theta_i) \rightarrow I_t. \quad (1.11)$$

On the other hand, the characteristics of an institutional arrangement at a given time affect how the culture of a group would evolve over time.

$$F_t(\theta_i) \xrightarrow{I_t} F_{t+1}(\theta_i). \quad (1.12)$$

Then again, the evolved culture affects the optimal design of a self-governing institution:

$$F_{t+1}(\theta_i) \rightarrow I_{t+1}. \quad (1.13)$$

The individuals in a social dilemma, then, not only have a chance to utilize a productive social opportunity, but they also have a chance to transform themselves into

a better group with wisely designed self-governing institutions. On the other hand, an institution designed with only a short-term time horizon and an overpessimistic view on the moral quality of the individuals involved can undermine a group's capability of self-governance in the long run. When everyone in a group is really selfish, it is unlikely that the group will successfully self-govern. Self-governance presupposes a certain moral quality of people. On the other hand, successful self-governance depends to a great extent on how it raises the moral quality of the self-governing people themselves.

1.3 The Plan of the Dissertation

Given the magnitude and complexity of the task of investigating motivation in social dilemmas within a game-theoretical framework, this dissertation can contribute only to a limited extent. In that regard, this dissertation focuses on a few critical aspects: game-theoretic modeling of heterogeneous motivations (multiple types), derivation of theoretical implications of the existence of multiple types to individual behavior and aggregate outcomes, and empirical tests of the game-theoretic analyses (Chapters 4 to 6).

As a groundwork for the theoretical and empirical work, the problems related to the definition of social dilemmas and the selfishness assumption (Chapter 2), and those related to the development of frameworks for the study of social dilemmas and formal models of heterogeneous motivations (Chapter 3) will be discussed.

Chapter 2 contributes to the groundwork for the investigation of non-selfish motivations. Some of the more fundamental issues need to be addressed when the universal selfishness assumption is discarded. The first question to be addressed in Chapter 2 is

whether to define a social dilemma in terms of von Neumann-Morgenstern utilities or observable material payoffs. Casual definitions that exist tend to obscure on this question, in spite that whether a social dilemma is defined in terms of observable material payoffs or in terms of utilities make a fundamental difference, especially in developing game theoretical models of a social dilemma. As a preparation, the concept of von Neumann-Morgenstern utility as a cardinalized preference is discussed. The chapter also revisits the selfishness assumption in economics and political science. This is important since, even after a social dilemma is defined in terms of observable material payoffs to facilitate empirical research, a casual introduction of the universal selfishness assumption prevents a modeling of heterogeneous motivations. The chapter reviews several justifications for the universal selfishness assumption and argues that there is no compelling methodological or empirical reasons to hold on to the long-held assumption, especially in the context of studying social dilemmas.

Chapter 3 develops a behavioral framework for the study of social dilemmas. A framework that incorporates the essential working parts of social dilemma action situations and their relationships is necessary to utilize the analytical power of non-cooperative game theory while addressing the problem of social dilemma defined by material payoff. The chapter critically reviews two such frameworks: (1) the Microeconomic System (MES) framework by Smith (1982) and (2) the Institutional Analysis and Development (IAD) framework by Ostrom, Gardner, and Walker (1994). Then, a behavioral framework for social dilemmas is developed incorporating the two frameworks of MES and IAD.

The framework serves as a bridge between the empirical world of social dilemmas and the mathematical world of game theory. Utility functions in the framework transforms

material payoffs (crucial empirical quantity in defining social dilemmas) into von Neumann-Morgenstern utilities (basic quantity of game theory). Arguments in the generic utility functions and alternative specifications of non-selfish utility functions are introduced in Chapter 3.

Chapter 4 develops a series of theoretical and empirical investigations of two alternative motivations to selfishness – altruism and inequity – in the context of single-play 2×2 social dilemma games. First, the two models' theoretical implications are derived with regard to preference ordering of players over the four possible outcomes of a 2×2 social dilemma, types of equilibria and supporting conditions, and the predicted relative frequencies of cooperative choices in the four qualitatively different information sets of the simultaneous and the sequential 2×2 social dilemma game. The implications are formulated as empirically testable hypotheses. Empirical tests of the hypotheses are conducted using two sets of experimental data. It will be shown that, while both the models based on inequity aversion and linear altruism account for the substantial proportion of behavior that cannot be explained when individuals are assumed to be entirely selfish, the inequity aversion model generates more clear and parsimonious explanations; the linear altruism model is too underdetermined.

Chapter 5 continues the analyses of Chapter 4 in the finitely repeated 2×2 social dilemma game setting. The sequential equilibria of the finitely repeated 2×2 social dilemma game are analyzed first, assuming that there are only two types of players and then assuming that there is a general distribution of types. The theoretical analysis of Chapter 4 provides an alternative to Kreps et al.'s (1982) model of cooperation in the finitely repeated Prisoner's

Dilemma. While Kreps et al. sacrifice the assumption of common knowledge inherent in modern game theory, the model developed in Chapter 5 does not. Based on the conclusions of Chapter 4, a series of realistic assumptions are made: (1) there is a significant proportion of individuals who are not purely selfish – who care about more than maximizing one’s own material wealth, (2) those non-selfish individuals tend to be conditional reciprocators rather than unconditional altruists, and (3) regardless of one’s motivation, individuals know that there exist non-selfish individuals. We will see that, in addition to an equilibrium in which all types of players defect in every stage, there exist equilibria in which cooperation occurs. In a cooperative equilibrium, two players cooperate from the first to near the final stage. In a hybrid equilibrium, players make a transition from mutual defection to mutual cooperation.

Chapter 6 conducts a series of empirical tests of the theoretical conclusions derived in Chapter 5 using an experimental data set. First, we will look at the outcome of each game and see how well the actual outcomes fit the types of equilibrium predicted in Chapter 5. All the types of equilibria are observed, including cooperative and hybrid equilibria. However, due to the existence of multiple equilibria, a majority of outcomes do not strictly follow a single equilibrium path. When stage game outcomes deviate from the equilibrium path, it is observed that players quickly coordinate on the path of one of the equilibria. Second, the impacts of theoretical variables predicted by the equilibrium analysis of Chapter 5 are tested in a series of regression models. Special issues related to the statistical analyses of experimental social dilemma data are addressed using the statistical models of bivariate probit and fixed effects logit. The fixed effects estimation reveals that individuals who

cooperate more in a phase do not necessarily cooperate at comparable levels in later phases. This is because, to play cooperation consistently, one not only needs to be a cooperative type, but also needs to meet a partner who is willing to play cooperatively. The bivariate probit analysis shows that players coordinate on an equilibrium path based on the way a game is played in previous stages.

Chapter 7 summarizes the theoretical and empirical findings of this dissertation and suggests possible directions to further develop this dissertation's research.

Chapter 2

Definition of Social Dilemmas and the Selfishness Assumption

2.1 Introduction

The goal of this chapter is to define the term “social dilemma” unambiguously. First, we will distinguish social dilemmas from the problem of collective goods provision. The essential characteristic of social dilemmas is the potential conflict between the interests of each individual and the group as a whole, which is not necessarily the case in the problems involving the provision of collective goods. Second, after social dilemmas are distinguished from other social problems, how to understand the concept of self-interest becomes the key question in reaching the final and unambiguous definition of social dilemmas.

Self-interest can be defined either in terms of von Neumann-Morgenstern utilities – the basic scientific quantity in game theory – or in terms of empirically observable material

rewards/payoffs structure. Whether self-interest is defined in terms of utilities or material rewards has far-reaching ramifications regarding what are social dilemmas and how to study them.

After critically reviewing the common definition and the utility-based definition of a social dilemma, this chapter proposes a behavioral definition of a social dilemma that is based on empirically observable material reward structures. A utility-based definition of a social dilemma hinders empirical research due to the unobservability of utilities. The nature of von Neumann–Morgenstern utility, and the inconsistencies in explaining cooperation when social dilemmas are defined by utility, will be discussed.

Finally, even after the concept of a social dilemma is defined in terms of the empirically observable material reward structures, the widely held assumption of self-interest that claims that individuals' preferences are based solely on their and only their material rewards in the alternative outcomes of an action situation, creates another key problem in explaining and facilitating cooperation in social dilemmas. Section 3 provides a critical review of the diverse ways in which the self-interest assumption has been justified. With the material reward-based definition of social dilemmas and the recognition of heterogeneity in individual motivations, Section 4 proposes the need for a behavioral framework for the study of social dilemmas.

2.2 Defining Social Dilemmas

2.2.1 Social Dilemmas and Collective Action Problems

The generic action situation to be investigated in this study is a *social dilemma*. “Collective action problem,” or “collective action situation” is often used interchangeably with “social dilemma.” Insofar as the concepts are defined clearly, the interchangeable usage is not problematic. However, there is also an interesting way to make a subtle and potentially productive distinction between the two concepts. From Olson (1965) to Hardin (1982), regardless how the authors defined a collective action problem, the actual studies have focused on the problem of providing collective goods. Conflict between individual and group interests, and resulting failure in providing collective benefits, are without doubt the most significant aspects in the provision of collective goods. In that sense, collective action problems and social dilemmas have been used interchangeably without any significant objection.

However, the conflict, even a potential one, between the group as a whole and each individual is not an inevitable aspect in the provision of collective goods. This is the case even when all the individuals are selfish and the provision problem is of a one-shot nature.

Provision of collective goods involves the conflict between the group and each individual only under certain circumstances. For example, when a privileged group or individual exists, a collective good can be provided even when every single individual acts in a short-term selfish manner. It is possible that the provision occurs at a suboptimal level leaving at least one other cost/benefit sharing arrangement in which everyone would be better-off. A more dramatic case is the provision of a collective good of which the provision level has a

ceiling. In that case, the collective good can be provided by a privileged individual without leaving any other Pareto superior cost/benefit allocation scheme. Laboratory experiments of linear public goods provision provide another example. When the marginal per capita return in such experiments is greater than 1, there is no conflict of interest between individual and the group. Therefore, those experimental situations are not social dilemmas. But it definitely is a case, if not problem, of a collective good provision.

The concept of “social dilemma,” on the other hand, directly focuses on the conflict between the interest of a group as a whole and that of each individual. Many social dilemmas do involve, or can be framed as involving, provision of collective goods. But there are also cases of social dilemmas that do not involve a collective good. The dilemma that two suspects face in the original story of the Prisoner’s Dilemma game is a good example. The prosecutor’s clever scheme intermingles the two suspects’ fate. But the prisoners are not involved in any problem of providing a collective good defined in economic theory, though they might be involved in a problem of providing a common good as the term is understood in political philosophy.

2.2.2 The Common Definition of Social Dilemma

It is generally recognized that a social dilemma (collective action problem) involves three key aspects:

- a group of individuals;
- a common interest among them; and
- a potential conflict between the common interest and each individual’s interest.

The behavioral consequence of the action situation, whether or not and to what extent the group of individuals achieves their common interest, is not an essential part of the definition of a social dilemma. However, certain types of hypothetical individual motivations and their implications for individual behavior and aggregate outcome are. So the definitions usually have the structure of “if individuals act selfishly ... they do not achieve a certain goal.” Even in this line of casual definition, the nature of the common goal and the action problem the individuals have are not always clearly stated.

For example, Dennis Chong (1991:5) says that a collective action “*problem*” arises “when individuals, acting out of pure self-interest, are unable to coordinate their efforts to produce and consume certain public goods they find desirable.” The context in which the above definition is provided leads the readers to think that what is proposed is a definition of a social dilemma (collective action problem) commonly understood. However, given that the problem of coordination is usually regarded as distinct from that of a social dilemma, the use of the term “coordinate” to characterize the desired collective action is not precise enough. In addition, “the provision and consumption of certain public goods” is not inclusive enough to denote the collective goal individuals may have in social dilemmas.

Olson (1965:2) provides a definition of a collective action problem that is closer to what this study calls the common definition of a social dilemma. According to Olson, in a collective action situation “... *rational self-interested individuals will not act to achieve their common or group interest.*” There are two ways to understand this definition. First, it can be understood as a combination of a definitive assumption about individuals – that they are rational and self-interested – and a strong prediction that they will *fail* to achieve

their common interest. Second, we can take the “rational self-interestedness” of individuals as a hypothetical statement and understand the definition as providing a logical consequence of the assumption given the characteristics of an action situation. In either case, the formal structure of the underlying action situation remains the same. But in the latter interpretation of Olson’s definition, the focus is the structure of the action situation while the individuals’ motivations and the success and failure of achieving the common interest are the subjects of research.

Ostrom (1998:1) provides a careful definition of social dilemmas that corresponds to the second interpretation of Olson’s definition of a collective action problem. “Social dilemmas occur whenever individuals in interdependent situations face choices in which the maximization of short-term self-interest yields outcomes leaving all participants worse off than feasible alternatives.” Below, a precise statement of the common definition of a social dilemma is provided in which the hypothetical nature of the motivational statement and the key factors involved in a social dilemma are stated explicitly.

Definition 1 (Common) *A social dilemma is an action situation involving a group of individuals and commonly valued good(s) where if each individual pursues his/her self-interest, the aggregate collective outcome(s) results in which each individual is worse off than he/she could be in at least one alternative feasible outcome.*

2.2.3 Definition of Social Dilemmas in Terms of Utilities

The common definition of a social dilemma is not satisfactory enough because the meaning of self-interest is not clearly defined. There are two alternative ways to define

self-interest: in terms of von Neumann-Morgenstern utilities and in terms of objective material payoffs. Obscurities found in the literature concerning this definitional problem have contributed significantly to the unnecessary confusions and debates over the substantive aspects of the issues related to social dilemmas.

In the utility-based definition, a social dilemma is formalized as a generic game with certain equilibrium characteristics.

Definition 2 (Utility-based) *A social dilemma is a game in which the equilibrium outcome(s) is(are) Pareto-inferior to at least one other outcome resulting from a non-equilibrium strategy profile.*

It is assumed that *reasonable* solution concepts of non-cooperative game theory will be used in calculating the equilibria. For example, in some sequential or repeated games with perfect information, the collectively optimal outcome could be reached by a Nash equilibrium. However, if the Nash equilibrium involves incredible threats or promises and the subgame-perfect equilibria meet the definition above, the game should be regarded as a case of social dilemma defined in terms of utilities. Some social dilemmas defined by utilities may have the dominant strategy equilibrium as is the case in the Prisoner's Dilemma. Others may not, for example, as in the standard Common-Pool Resources game with a decreasing marginal return to variable inputs.

The significance of social dilemmas lies in the very fact that they are everywhere in our daily life. To ultimately synthesize the massive empirical research on social dilemmas, the object of study itself needs to be defined in terms of empirically observable elements. In most of the cases, the material payoffs are commonly valued goods whose preservation,

production, and allocation are of utmost social concern.

On the other hand, the von Neumann-Morgenstern utility is a purely mathematical quantity that is not easy to measure.¹ This does not imply that non-cooperative game theory, of which the von Neuman-Morgenstern utility is the key scientific quantity, is not useful in studying social dilemmas. Quite to the contrary, this dissertation relies heavily on non-cooperative game theoretic analysis of social dilemmas. It is a contention of this dissertation that game theory needs to be used in developing models of empirically defined objects, but not to define the object itself.

The following two sections discuss the concept of von Neumann-Morgenstern utility and the problems arising in attempting to explain observed cooperative behavior while defining social dilemmas in terms of utilities.

2.2.4 The Nature of von Neumann-Morgenstern Utility

Von Neumann-Morgenstern utility is a scientific quantity that expresses preferences cardinally. As is well explained in Luce and Raiffa (1957:31-32), preferences come first and the concept of von Neumann-Morgenstern utility represents preferences in a cardinal way. Assume that an individual i strictly prefers option X to option Y . Social choice theory, in which cardinal values are usually unnecessary, has a simple way to denote this: xPy . Preference orderings over more than two objects can be represented in the same way. For example, “ x is preferred to y and y is preferred to z ” can be denoted as $xPyPz$. How-

¹Arguments can be made that there are techniques to measure von Neumann-Morgenstern utilities, especially in the laboratory setting. One way is to directly apply the lottery method that will be introduced in the next section. However, the method involves human choices that are stochastic in nature. Therefore, the most favorable statement about the measurement of von Neumann-Morgenstern utility is that it could be measured stochastically under very tightly controlled settings.

	X	Y
1	2	1
2	2	1.999
3	2	-100
4	0.2	0.1

Table 2.1: Cardinalization of preference over two events: xPy

	X	Y	Z
1	2	1	0
2	20	19	0
3	200	199	0

Table 2.2: Cardinalization of preference over three events: $xPyPz$

ever, one might want to know how strongly individual i prefers X to Y and Y to Z . One could ask “does individual i prefer X to Y more than he prefers Y to Z ? And how much?”

In von Neumann-Morgenstern utility theory, numbers are assigned to events, with a larger number assigned to an event representing that the event is more preferred to other events with smaller numbers assigned. If there are only two events over which the preference relationship is to be established, any two different numbers would work. All the sets of assigned numbers in Table 2.1 equally well represent a preference ordering xPy .

Now assume that there are three events, X, Y , and Z , and the individual i prefers X only slightly more than Y , but he prefers Y very strongly to Z . What are the three proper numbers to denote this relationship?

In Table 2.2, the order of the numbers assigned to the three events X, Y , and Z , in each of the three rows, equally well represents the preference ordering $xPyPz$. But since the numbers are equally distanced in the first row, one might not agree that the first case well represents the fact that the individual in our example prefers X to Y more strongly than he prefers Y to Z . Row 2 and row 3 satisfy the original intention better than row 1. But

which one is better? Or, is there any objective way to solve the problem? Since ordinality is the only natural content in preference relationships, cardinalization has to be indirect, in the sense that it requires another process by which the cardinalization can be obtained.² Von Neumann and Morgenstern cardinalize the preference relationship by combining events with probabilities.

Let us assign 0 to event Z and 1 to event X . Now the individual has to establish his preference between two options: option 1 gives him Y for sure, or with probability 1, and option 2 gives him a lottery in which he has a probability p of receiving X and a probability $1 - p$ of receiving Z . Let us denote these two options Y and $p(X, Z)$. In the von Neumann-Morgenstern utility system, the value of p that makes the individual indifferent between Y and $p(X, Z)$ is the numerical value that has to be assigned to Y . (In this example, we assumed a preference ordering $xPyPz$ and assigned 0 and 1 to the events X and Z). The resulting value of p depends on the numbers originally assigned to X and Z . That means that von Neumann-Morgenstern utilities are defined only up to linear transformation. In other words, like the concept of distance, there is no absolute unit for utilities.

Now it is possible to fully describe the substantive meaning of the utilities assigned to three events. Define $u(X)$ as the von Neumann-Morgenstern utility assigned to X . If $u(X) = 1$, $u(Y) = 0.3$ and $u(Z) = 0$ for an individual, it is meant that she prefers X to Y and Y to Z , and she is indifferent between Y and a lottery of $0.3(X, Z)$ - a lottery that gives her a probability 0.3 of receiving X and 0.7 of receiving Z .

The above procedure can be generalized to cases involving N events, X_1, X_2, \dots

²Temperature is an indirect cardinalization of the degree of warmth relying on the height of mercury contained in a glass bar.

X_i, \dots, X_N , with preference ordering $x_1 P x_2 P \dots x_i P \dots P x_n$ and the original numbers assigned to the most and the least preferred events are other than 1 and 0. First, assign any two different numbers $u(X_1)$ and $u(X_N)$ [$u(X_1) > u(X_N)$] to the most-preferred and the least-preferred events. Second, find the probability p_i to each of the events in between, $X_2, \dots, X_i, \dots, X_{n-1}$, that makes the individual indifferent between X_i and a lottery $p_i(X_1, X_n)$ in which individual i has probability p_i of receiving X_1 and probability $1 - p_i$ of receiving X_n . Then,

$$u(X_i) = p_i \times u(X_1) + (1 - p_i) \times u(X_N). \quad (2.1)$$

For example, suppose a case in which there are four possible events, X_1, X_2, X_3 , and X_4 , and your preference ordering over them is $x_1 P x_2 P x_3 P x_4$. Assign two arbitrary numbers 10 and 5 to the most and the least-preferred events X_1 and X_4 . By this we are setting $u(X_1) = 10$ and $u(X_4) = 5$. The next step is to find p_2 and p_3 that make you indifferent between X_2 and $p_2(X_1, X_4)$ and between X_3 and $p_3(X_1, X_4)$. Suppose $p_2 = 0.6$ and $p_3 = 0.3$. Then,

$$u(X_2) = 0.6 \times u(X_1) + 0.4 \times u(X_4) = 8 \quad (2.2)$$

and

$$u(X_3) = 0.3 \times u(X_1) + 0.7 \times u(X_4) = 6.5. \quad (2.3)$$

One does not need to assign the original two arbitrary numbers to the most- and

the least-preferred events to find the lottery probabilities for the events in between. In fact, the crucial element is the lottery probabilities for the events in between that make the individual indifferent. So the procedure described above can start with finding the probability first and then assign two arbitrary numbers to the most- and the least-preferred event, and finally calculate utility of the event in between the most- and the least-preferred events. In addition, one does not need to assign two numbers to the most- and the least-preferred event. One could start by assigning two different numbers to any two events in the event set.

One consequence of this definition of utility is that utilities are defined only up to linear transformation. Thus, transformation of $u(X)$, $u(Y)$, and $u(Z)$ into $a \times u(X) + b$, $a \times u(Y) + b$, and $a \times u(Z) + b$, (where a is greater than zero) does not alter the original substance. By definition, utility represents individual preference. Therefore, interindividual comparison is not possible in the von Neumann-Morgenstern utility system. Payoffs of a game in standard game theory are von Neumann-Morgenstern utilities defined as such. Figure 2.1 shows three game matrices in which the payoffs are von Neumann-Morgenstern utilities. What are the differences among these three games?

Insofar as the payoffs of the three games in Figure 2.1 are von Neumann-Morgenstern utilities, it has to be clear to the readers that the three games are identical. Two basic features of the von Neumann-Morgenstern utilities make the three games identical. First, utilities are defined up to linear transformation. Second, utilities are defined for each individual independently; thus, interpersonal comparison of utilities is not possible.

Strictly speaking, a perfect representation of a naturally occurring action situation

		<i>Player 2</i>	
		<i>L</i>	<i>R</i>
<i>Player 1</i>	<i>U</i>	1, 1	0.2, 0.6
	<i>D</i>	0.6, 0.2	0.8, 0.8

		<i>Player 2</i>	
		<i>L</i>	<i>R</i>
<i>Player 1</i>	<i>U</i>	10, 10	2, 6
	<i>D</i>	6, 2	8, 8

		<i>Player 2</i>	
		<i>L</i>	<i>R</i>
<i>Player 1</i>	<i>U</i>	1, 10	0.2, 6
	<i>D</i>	0.6, 2	0.8, 8

Figure 2.1: The Same Game?

as a game involves not only the knowledge of all individuals' ordinal preference orderings over all the possible outcomes but also each individual's preference ordering over any outcome and lotteries involving other outcomes.

There are purely game-theoretic, but still interesting and important issues, related to social dilemmas defined in terms of utilities. But a serious problem exists in empirical research. A researcher usually does not know exactly the motivations of individuals in an action situation. Therefore, in most of the cases, it is the researcher's assumption about individual motivations that results in the characterization of an action situation as a social dilemma defined in terms of utilities. Or, if the research has in-depth knowledge about individuals motivations, many action situations commonly understood as social dilemmas (organization of a labor union, provision of public goods, etc.) will have to be excluded from the category of a social dilemma.

2.2.5 Standard Non-cooperative Game Theoretical Approach to Social Dilemmas

This section discusses the inconsistencies in the attempted explanations of observed cooperation in social dilemmas defined in terms of utilities. The *Prisoner's Dilemma* game is the most famous and simplest case of a social dilemma defined in terms of utilities. In Figure 2.2, $T, R, P,$ and S are von Neumann-Morgenstern utilities and a relationship holds among them: $T > R > P > S$. Regardless of what the other player chooses, a player always prefers the outcome resulting when he himself chooses D . Therefore, both players are said to have a dominant strategy.³ The unique equilibrium of this game is (D, D) and corresponding payoffs to players (P, P) . But since R is greater than P , (D, D) is strictly Pareto-inferior to (C, C) . In studying single-play *Prisoner's Dilemma* games, the exact cardinal values of payoffs are not really important insofar as the ordering among the four outcomes hold.

If a social dilemma is defined in terms of von Neumann-Morgenstern utilities, is there any way to explain cooperation? Put another way, can *Cooperation* occur in equilibrium of the *Prisoner's Dilemma* game, the simplest social dilemma game defined in terms of utilities? First, it is well-known that *Cooperation* in single-play or finitely repeated *Prisoner's Dilemma* cannot be a part of an equilibrium. Only when the game is repeated infinitely, do there exist equilibria other than a sustained defection by all the individuals. Probabilistic continuation of the game only opens the possibility for achieving

³There are also social dilemma games in which players do not have dominant strategies, but equilibria are still Pareto inferior. The standard common-pool resource game is an example (see Walker, Gardner, and Ostrom, 1990).

better collective outcomes than that in a single-play or finitely repeated game.

But what about the well-known theories of cooperation in the finitely repeated *Prisoner's Dilemma* games (Kreps et al., 1982; Fudenberg and Maskin, 1986)? The remainder of this subsection carefully follows the logic of Kreps et al. to show that cooperation is still not possible in *the finitely repeated Prisoner's Dilemma game defined in terms of utilities*.

In Kreps et al., either at least one of the essential assumptions of non-cooperative game theory is relaxed or the payoffs of the game are understood as material rewards to reach the possibility of cooperation. Relaxation of the formal requirements of game theoretic analyses is problematic because anything can be shown to be possible that way. *De facto* interpretation of payoffs as material rewards is another analytical strategy with which Kreps et al. show the possibility of *Cooperation*. In that case, what Kreps et al. analyze is the simplest case of a social dilemma defined in terms of material payoffs, not the *Prisoner's Dilemma* game.

Kreps et al. provide two models of cooperation in the finitely repeated games:

1. When a rational player has a small prior that her partner is irrational, in the sense that the partner plays *TFT (Tit-For-Tat)* strategy, the rational player has an incentive to cooperate in earlier stages of the finitely repeated game.
2. If there is a common prior that a player enjoys *mutual Cooperation*, then *Cooperation* by both of the players until near the end of the game is an equilibrium.

There is more than one way to interpret the *TFT* strategy in Kreps et al.'s model. First, as the authors sometimes say, the *TFT* strategy can be understood as a result of a

player's irrationality. Then again, there are three ways to interpret the meaning of irrationality. First interpretation of irrationality is that the utility function cannot be defined for this kind of players. However, this is a violation of one of the most important assumptions of the standard non-cooperative game theory. Strictly speaking, when utility function cannot be defined for players, the solution concepts of non-cooperative game theory cannot be applied.⁴

Second, it is also possible to understand a player's irrationality in the sense that there is a payoff function for him and utilities can be defined, but that he does not follow the expected utility maximizing strategy. But this interpretation contains a potential paradox. Or, this interpretation introduces an ontological dilemma related to defining preference which may not be materialized in action.

The third way to interpret the existence of irrational players is that they do not actually need to exist. All that is necessary is some rational players' subjective prior that others might be playing *TFT*. This interpretation, while not being a violation of the rationality condition, does violate another fundamental assumption of the non-cooperative game theory: the common knowledge condition. In the standard non-cooperative game theory, the three basic elements of a normal form game – players, strategies, and payoff function – have to be common knowledge among all the players. This assumption is necessary to avoid possible arbitrariness when different priors are allowed.

Any outcome from any game can be rationalized once it is allowed that players understand the game differently. The common knowledge condition does not require perfect

⁴As a matter of fact, what Kreps et al. call equilibrium of their first model is a profile of irrational player's presupposed action and rational player's solution to her decision-making problem.

	COL	
ROW	<i>Fink</i>	<i>Cooperate</i>
<i>Fink</i>	0,0	a, b
<i>Cooperate</i>	b, a	1,1

* $a > 1$, $b < 0$, and $a + b < 2$

Figure 2.2: Stage Game Matrix of the Finitely Repeated Prisoner's Dilemma Game. Source: Kreps et al.(1982:245).

information. Harsanyi (1967-68) has proposed a way to achieve common knowledge when information asymmetries exist among players. The essence of the "Harsanyi transformation" is that players, even when they do not exactly know the payoff function of the other players in a game, do share a common belief regarding the probability distribution of player types defined in terms of the payoff function. For Kreps et al.'s model to meet the common knowledge condition, there should be a commonly known (objective) probability that a player is a *TFT* type. But if there really exist some *TFT* players, then we fall back to the problem of how to understand their irrationality.

The second model of Kreps et al. posits a small objective probability that one's "opponent 'enjoys' cooperation when it is met by cooperation" (251).⁵ The repeated game analyzed in Kreps et al. consists of " N repetitions of the ... two person, bimatrix, stage game" shown in Figure 2.2. The "enjoy-cooperation" model posits a small probability δ that $a < 1$.

First, if the entries of the original matrix represent utilities, then the meaning of "enjoying cooperation" becomes obscure. In principle, the payoffs of a game should represent all the factors that affect preference. Second, if the entries of the original matrix

⁵Fudenberg and Maskin (1986) express this idea as a type of players having a very complicated payoff function that makes *TFT* a rational strategy for them.

represent only material rewards, the utilities can be given only after the “enjoyment” factor is taken into account and, thus, the entries in Figure 2.2 are not payoffs of the game. In either of the two interpretations, if there exists a proportion of players who enjoy mutual cooperation, the basic *Prisoner’s Dilemma* game of Figure 2.2 is not the stage game of the repeated game analyzed by Kreps et al. Or, to put it differently, the repeated game analyzed in Kreps et al. has a stage game that is different from the *Prisoner’s Dilemma* shown in Figure 2.2. Therefore, their analysis is not that of the repeated *Prisoner’s Dilemma* game.

In sum, Kreps et al.’s models do not provide a satisfactory game theoretic explanation of *Cooperation* in social dilemmas because the object of study is not clearly defined and/or the formal requirements of game theoretical analyses are often violated.

2.2.6 A Behavioral Definition of Social Dilemma

Each possible interpretation of Kreps et al.’s models is at the same time a way to explain observed cooperation in the laboratory *Prisoner’s Dilemma* games. Some players may not have fixed preferences defined over the four outcomes of the stage game and just follow the behavioral rule of *TFT*. Or, individuals may be all rational and selfish, but do not share a common prior that all of them are rational egoists. Or, maybe there exist individuals who are rational but not selfish and their existence is common knowledge. While the first and second interpretations cannot be easily discarded, they require analytical tools other than game theory.

The last interpretation of Kreps et al.’s model is the most promising to develop a game theoretic model that explains the observed cooperative behavior in the laboratory *Prisoner’s Dilemma* game. It enables one to define the action situation based on empir-

ically observable factors and provides a way to conduct game theoretical analyses without abandoning the rationality and common knowledge conditions.

However, the meaning of “enjoying of cooperation” has to be further developed into a more rigorous proposition. Do some individuals enjoy the *action* of cooperating itself? Or, do they enjoy the *outcome* of mutually cooperative choices? Do some individuals care for others’ well-being? Are they concerned with fairness/reciprocity? The “enjoyment” assumption in Kreps et al. was only a technical convenience to make the stage game somewhat different from the strict *Prisoner’s Dilemma*. To develop an empirically valid theory of cooperation in social dilemmas, the nature of non-selfish motivations needs to be addressed in a more rigorous way.

This emphasis on material objects does not lead to crude material understanding of human interactions. Quite to the contrary, it is only after a generic action situation is clearly defined based on the empirically observable factors, that an in-depth discussion of the ethical and cultural factors can start.

Material payoffs are measures of goods that individuals receive as a function of their own and others’ choices. Money is, without doubt, a measure of material payoffs. Most experimental research, and many naturally occurring action situations, directly involve allocation and distribution of money. Or, in the case of the two suspects in the original tale of the Prisoner’s Dilemma game, the material payoff is the liberty, the length of free time enjoyable out of the prison.

Definition 3 (material reward-based 1) *A social dilemma is an action situation involving a group of individuals and commonly valued good(s) in which if each individual*

tries to maximize his/her material reward, the outcome(s) results in which each individual is worse off (measured in their material well-being) than he/she could be in at least one alternative feasible outcome.

It is worth re-emphasizing that the “maximization of material reward” is a hypothetical statement to clarify the reward structure of the action situation. If each individual indeed tries to maximize his/her material reward, the game representation of a social dilemma would be the same as the social dilemma defined in terms of utilities.

The choice of action that maximizes one’s material reward may be unconditional as is the case, for example, in the experimental linear public good provision problem with the marginal per capita return of less than 1. Or, it may be conditional and depend on other individuals’ choices. In any case, the material reward maximizing behavior needs to be analyzed borrowing from the game theoretical framework. First, assume that for all individuals, the utility from each outcome of the action situation is a one-to-one mapping of one’s own material reward. Second, calculate the reasonable equilibria of the game constructed with the motivational assumption. Third, conclude that the action situation is a social dilemma if the equilibria of the game constructed as such are Pareto-inferior to the outcome resulting from at least one non-equilibrium strategy profile.

The game constructed as such might have multiple equilibria. If all of them are Pareto-inferior to at least one non-equilibrium outcome, the original action situation is a social dilemma. If at least one equilibrium gives an outcome on the Pareto-frontier, the underlying action situation is not a social dilemma.

Notice that here the game theoretic discussion of the action situation is to clarify

the material reward structure with hypothetical and, probably, unrealistic, motivational assumption. The more important game theoretic analyses should be done after heterogeneous motivations are incorporated into the material reward structure of the social dilemma action situations.

We have consciously avoided using the term *self-interest* in the material-based definition of social dilemma. In this study, an individual is said to be self-interested if the sole purpose of his/her action is to maximize the material reward he/she gets. The material-based definition of social dilemma can be rewritten using the self-interest concept.

Definition 4 (material reward-based 2) *A social dilemma is an action situation involving a group of individuals and commonly valued good(s) where if each individual acts on his/her own self-interest, an outcome results in which each individual is worse off than he/she could be in at least one alternative feasible outcome.*

In explaining individuals' behavior in social dilemmas, understanding the heterogeneous motivations is most important. It is still possible to model heterogeneous motivations while defining self-interest as utility maximization. However, in that case, all the non-material maximizing patterns of behavior fall in the rubric of self-interested behavior. A very awkward classification of motivations will be inevitable in which behavior that sacrifices one's own well-being for the sake of others', that follows certain ethical rules, etc., are all called self-interested. Defining self-interestedness as material reward maximization makes the discussion regarding heterogeneous motivations more productive.

2.3 A Critical Review of the Selfishness Assumption

The “if ... then” structure of the material-based definition of a social dilemma highlights the structural characteristics of the action situations. Explaining cooperative behavior, then, takes two forms. First, one can assume that individuals are in fact self-interested and explain cooperative, thus apparently anomalous, behavior, in terms of the changes in the structure of the action situation. The state, market, and self-governing solutions reviewed in Chapter 1, when understood as empirical explanations, are the examples. In fact, the overwhelming majority of empirical and theoretical investigations thus far have followed this approach.

Alternatively, one can acknowledge the existence of non-selfish motivations and model them explicitly. This approach does not deny the importance of the non-motivational factors in social dilemmas. To the contrary, the purpose of this approach is to incorporate a well-established fact that there exists rational non-selfish individuals and synthesize both the motivational and non-motivational factors in a theoretically coherent and empirically valid manner.

Why has the self-interest assumption become so popular in studying human social interactions? What justifications have existed to defend it against the apparent anomalies? A contention that human beings are all selfish (in the sense of material well-being maximization) sounds too strong. Milton Friedman (1954), thus, had to defend the assumption in a rather distorted way. His defense of using a universal selfishness assumption in building theory did not rely on the assumption’s empirical validity. However, to consider the conformity of a theory’s assumptions to reality as a test of its validity is, Friedman argues,

fundamentally wrong (14). Assumptions can never be realistic. The only proper test of a theory is whether or not its implications survive empirical tests.

Whether assumptions are free from any direct empirical scrutiny is not clear in Friedman's argument. At some point, using a biological example of photosynthesis, he seems to argue that hypotheses are completely free from reality tests; insofar as the hypothesis that "leaves deliberately sought to maximize the amount of sunlight it receives" explains "the density of leaves around a tree" (19) well, the hypothesis is justifiable. On the other hand, he also suggests that hypotheses cannot be "descriptively" realistic since they are inevitable abstractions from the reality. But hypotheses, rather than being arbitrary, need to abstract "the common and crucial elements from the mass of complex and detailed circumstances" to explain the object of the theory in the most efficient manner.

Thus in the study of market economies, the hypothesis that "individual firms behave as if they were seeking rationally to maximize their expected returns and had full knowledge of the data needed to succeed in this attempt" is valid insofar as it abstracts the essential aspects of the reality efficiently and generates empirically valid predictions.

Demsetz (1997) takes a more straightforward position regarding the selfishness assumption in *economics*. Demsetz argues that abstraction and reality are not mutual enemies. Abstraction is necessary in that complete description yields not theory but only classification. However, pure abstraction itself cannot advance empirical science since it only produces mathematical logic. Therefore, what is important is the manner in which the abstraction is conceived. In that regard, "the neoclassical conceptualization, contrary to what its critics say, is a *realistic* portrayal of those characteristics that are important

in an inquiry into the *commercial activities of large, decentralized economic systems*"(4). Notice that Demsetz restricts the domain to which the assumption is a realistic abstraction: inquiry into commercial activities of large, decentralized economic systems.

Alchian (1950) provided a stronger case for self-interest/rationality assumption for economics based on the widely accepted theory of the market. The evolutionary process operating within the market drives out non-profit-maximizing, non-calculating actors provided that the competition among firms is sufficiently intense. Therefore, even when the original actors themselves do not consciously pursue those goals in that manner, it is exactly those actors with the characteristics consistent with economic theory who survive in the market.

This perspective has been revived in recent debates among political scientists over issues related to rational choice theory. Satz and Ferejohn (1994) argue that the self-interest as an empirical foundation of rational choice theory is not an individual-level psychological attribute, but a consequence of structural characteristics of certain arenas of human interaction. Their argument is based, similar to that of Alchian, on the paradigm of evolutionary biology that views nature as "selective structure"(81). Therefore, rational choice theory is most successful where "individual action is severely constrained, and thus where the theory gets its explanatory power from structure-generated interests and not from actual individual psychology"(72).

It becomes rather puzzling, given the theoretical background of the self-interest assumption in economics, why the rational choice theorists in political science have relied so much on the self-interest assumption in their study of collective action/social dilemmas.

In modern political thought, the original emphasis has been laid more on “self” than “interest” part of “self-interest.” It was the seventeenth century contractarians including Hobbes and Locke, who provided foundations for modern democratic ideas based on self-interest. In spite of the practical conclusions that often justified existing monarchy, the true revolutionary aspect of the political philosophy in that era was contained in the kind of thinking that “every one by the light of nature and reason will do that which makes for his greatest advantage” (Lee, 1978[1656] – quoted in Mansbridge, 1990:5).

In the twentieth century, Schumpeter (1942) went even farther to challenge the very notion of common good in defining the meaning of democracy. He laid out a view of politics and democracy in which competition among politicians is described as analogous to that among firms for market share. The relationship between political leaders and citizen voters is compared to the relationship between firms and consumers.

This perspective, often called “adversarial democracy,” still leaves room to interpret the meaning of self-interest in many different ways. In other words, self-interest in politics need not necessarily be equated to the pursuit of immediate material gain. As a founding block of modern democratic politics, self-interest could simply mean the pursuit of what one prefers without restricting the content of what a citizen might prefer. In addition, often in the realm of politics, the short-term, as well as the long-term, wealth effect of important public issues to each citizen is not clear.

However, in part due to the spread of analytical methods developed in economics and in another part due to the influx of new subjects of political inquiry, the narrowest sense of self-interest has come to occupy an increasingly important role in political analysis.

Two major themes in Downs's seminal book *An Economic Theory of Democracy* (1957) exemplify the transition. First, in what later became the spatial theory of voting, each citizen has preferences over issues, and politicians choose positions attempting to maximize the number of votes they receive in an election. On the part of politicians, the meaning of self-interest can be rather narrowly defined: the pursuit of an immediate material object, the vote. On the part of citizens, what is necessary is their preferences over issues. The motivational foundations for such preferences need not be limited to self-interest. A citizen's position on an issue, for example tax policy, could be based on the amount of material gain she expects from a certain policy content. But whether a citizen's position on tax policy is based on her material gain or her sense of social justice is not an essential part of spatial theory of voting. The position itself, and its distribution among a constituent, is what matters analytically.

In the other part of *An Economic Theory of Democracy*, Downs creates a major puzzle by using the notion of self-interest in its narrowest sense. The puzzle is citizens' participation in voting. The small material costs involved in voting participation in the form of time forgone and transportation cost exceed the expected gains from casting a vote given the extremely small probability that one's own vote becomes pivotal.

The increasing use of the self-interest assumption defined as the pursuit of one's own material gain can also be associated with the proliferation of applied game theoretic models of collective action. Game theory itself does not require the self-interest assumption. But when empirical questions are addressed in a game theoretical framework, it is often necessary and convenient to adopt the self-interest assumption. The reason is that the

match between the core quantity of game theory – utility – and the core empirical quantity in social dilemmas – material rewards – make the task of building game theoretical models of collective action easier.

Increasing numbers of political scientists and economists conducted empirical and theoretical research on collective action using game theoretic models that assume the correspondence between material rewards and utilities. Olson's *The Logic of Collective Action* (1965) was a pivotal work in presenting the idea of collective action problems at the general level. In terms of substance, Olson relied heavily on the example of labor movement/union participation. Again, the dilemma was that in spite of the fact that every potential member of a labor union prefers to have the benefits the union can provide for him, the collective nature of the benefit makes him free-ride on the others' contribution in the establishment and operation of the labor union. This, of course, is not a puzzle at all, for those who disagree with the fundamental assumption in it; that every worker is self-interested. Many puzzles related to social dilemmas are proposed and found in this manner.

Based on this short review of the relevant history, it is time to reflect whether there is any compelling reason to preserve the selfishness assumption in the study of social dilemmas. Friedmanian instrumental justification that a theory should not be judged by the empirical validity of its assumptions would no longer hold, given the vast amount of anomalies in the theories of collective action based on the assumption of self-interest. The assumption does not seem to be a realistic abstraction in Demsetz's sense in the areas of social dilemmas, though it might be the case in the study of competitive markets. This leads to a rejection of the justification that views selfishness as an innate psychological tendency

of human beings. In addition, the evolutionary (Alchian) and structural (Satz and Ferejohn) defenses of the assumption are not persuasive when applied to social dilemmas, since being selfish is not the best way for individuals to achieve the most out of social dilemmas.

2.4 The Need for a Behavioral Framework

Empirically oriented scholars of politics have been well aware of the fact that not every individual is selfish. Various kinds of non-selfish deeds are very common in politics among leaders and citizens. Then, why have the rational choice theorists of collective action not been able to dispense with the selfishness assumption?

There seem to be two major reasons. First, while borrowing the method developed in economics, many political scientists also borrowed the self-interest assumption without noticing that the various justifications for the assumption in the study of market economies do not hold quite well in the study of social dilemmas. Second, there are practical difficulties involved in building formal models of social dilemmas without the strong assumption of self-interest.

Many political scientists who are not antagonistic to the rational choice approach have pointed out the empirical anomalies arising when the self-interest assumption is adopted (see articles in Mansbridge, 1990). Some attempts have also been made to develop formal models that do not rely on the assumption of self-interest. However, a full-blown, as well as empirically relevant, game theoretical analysis of social dilemmas has remained a difficult task.

What does it mean that not every individual is selfish? It is only an argument

against a very strong and simple idea that *everyone is selfish*. Refuting the selfishness assumption opens up an infinite number of modeling possibilities. Game theoretic analysis of the social dilemma needs an assumption about players' motivation that is expressed in terms of their utility function that transforms the material payoffs into von Neumann-Morgenstern utilities. The strong self-interest assumption corresponds to a utility function

$$u^i = u^i(x^i) = x^i \quad (2.4)$$

where x^i is the material payoff for an individual i in the action situation.⁶ Then any utility function other than (2.4) would qualify as representing non-selfish motivations. The first difficulty in game theoretic modeling of social dilemmas without the selfishness assumption is that there are too many alternatives. At this stage, all we can say about the alternative non-selfish utility function is that its generic form has arguments other than one's own material payoff.

$$u^i = u^i(x^i, Z) \quad (2.5)$$

where Z is a vector of any elements other than one's own material payoff.

The self-interest assumption also implies homogeneity among individuals: that *everyone* is selfish and everyone is the *same* in that regard. Common sense suggests that individuals are different from each other. Facing social dilemmas, some would still behave selfishly while others would not. Those who are not strictly selfish would differ among

⁶This utility function also assumes that individuals are risk neutral. A selfish utility function without any assumptions regarding risk attitude would be : $u^i = u^i(x^i)$, where $\frac{du^i}{dx^i} > 0$.

themselves in terms of how far they depart from the purely selfish motivation. In game theoretic analysis, this necessitates the introduction of multiple types. A generic utility function, then, needs to include type parameters as well as arguments other than one's own material payoff:

$$u^i = u^i(x^i, Z : \Theta^i) \quad (2.6)$$

where Θ^i is a vector of individual i 's type parameters.

Nearly half a century of experimental research on social dilemmas has provided an enormous amount of empirical evidence that the selfishness assumption expressed in utility function (2.4) does not go far in explaining individuals' behavior outside of market environments. At the same time, the exploration of alternative motivations and its incorporation in game theoretic framework has only recently been launched (Crawford and Ostrom, 1995; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Cain, 1998; Daugherty and Cain, 1999; Rabin, 1993; Falk and Fischbacher, 1998; Dufwenberg and Kirchsteiger, 1998). Rigorous game theoretic analyses based on the recognition of multiple types and uncertainty are rare in spite of the methodological tool already provided by Harsanyi (1967-68) a few decades ago.

Final preparation to develop an empirically grounded theory of cooperation in social dilemmas is to develop a behavioral framework. The framework should provide the linkage between the two research strategies so far established for this dissertation.

1. The object of study – social dilemmas – needs to be defined based on the observable factors in the action situation.
2. Game theoretical analysis needs to be done while acknowledging the existence of

multiple types of individuals.

A framework, first of all, identifies essential working parts of the empirical world. Game theory itself is a mathematical tool that does not address the empirical world directly. Crucial working parts of social dilemma action situations, including material payoffs but not limited to them, need to be identified based on a framework of social dilemmas. The framework at the same time establishes the static and dynamic relationships among the working parts of an empirical world. The framework will be especially helpful in formalizing the evolutionary dimension of a social dilemma.

Finally, the framework has to relate its empirical working parts to the theoretical working parts of game theory. This will help in understanding the empirical implications of game theoretical analyses. At the same time, since not every empirical working part in a framework has its place in the game theoretical analysis, the framework will make it clear what assumptions are made and how they are justified in game theoretical analyses of an essentially empirical object. Chapter 3 develops a behavioral framework for the study of social dilemmas and explains its basic operations.

Chapter 3

Frameworks and Models

3.1 Introduction

A framework is a configuration of “the broad working parts and their posited relationships that are used in an entire approach to a set of questions” (Ostrom, Gardner, and Walker, 1994:25). A framework relevant for the study of social dilemmas needs to abstract the crucial working parts involved in the generic action situation of social dilemmas. Inductive theories of social dilemmas establish empirical regularities within the subsets of the working parts of the framework. For example, “communication among individuals has a positive impact on achieving a higher level of cooperation” and “the larger the gain from mutual cooperation, the more likely that an individual will cooperate” are empirical theories of social dilemmas. On the other hand, game theory and decision-making theories are deductive theories that can be used within the broader framework of social dilemmas to derive hypotheses regarding the relationships among working parts or to generate models of specific social dilemma action situations.

A framework itself does not yield empirical arguments. On the contrary, a better framework is one that is compatible with multiple theories, that helps a researcher to understand which theory is more suitable in explaining which kind of social dilemma action situation. The overall advancement in the study of social dilemmas can be made through the dynamic interaction between deductive and inductive theories within a well-articulated framework. The results of advanced knowledge of social dilemmas accumulate as empirical theories of social dilemmas.

The primary purpose of this chapter is to develop a *behavioral framework* of social dilemmas. The term “behavioral” implies that the framework is intended to be useful in explaining empirical phenomena and not just a meta-theoretical device.¹ Given the definition of a framework described above, the term behavioral is rather redundant. However, the redundancy helps to emphasize the need to develop a framework that allows a rigorous empirical approach to analyzing social dilemmas. On the other hand, the term “behavioral” is added to indicate an extensive use of “behavioral game theory” in conjunction with the framework in this study.

Behavioral game theory itself has a far broader scope of interest. Camerer (1997) defines behavioral game theory as an approach that “aims to describe actual behavior.” Camerer identifies three main issue areas where standard game theory has experienced frequent failure and behavioral game theory tries to provide more plausible explanations:

¹However, it is different from the Skinnerian sense of behaviorism, which is a particular case of empiricism. In the Skinnerian behaviorism, the primary focus of research is the establishment of empirical regularities between the kinds of stimuli an organism receives and the kinds of reactions by the organism to the stimuli (Skinner, 1974). Articulation of the inner mechanisms that connect stimuli and responses is intentionally avoided in Skinnerian behaviorism. Since the *behavioral framework* of social dilemmas does intend to articulate the inner mechanisms of individuals in the forms of preference and information processing, it is on the line of neo-behaviorism in the psychological discipline.

(1) social preference, (2) problems of choice and judgement that include framing effects and players' overestimation of their own capabilities, and (3) problems of strategic reasoning that include focal points and timing of the choice.

Behavioral game theory is still a branch of (non-cooperative) game theory in that it relies on the solution concepts of non-cooperative game theory to link the initial conditions of a game and the resulting behavior and aggregate social outcome. To describe and explain actual behavior while preserving the essential theoretical power of game theory, behavioral game theory needs a framework – a system of formal language – that can capture what standard game theory misses, while at the same time formal and logical enough to utilize the analytical concepts of noncooperative game theory.

Section 2 reviews two exemplary frameworks: V. Smith's Microeconomic System (MES) framework and the Institutional Analysis and Development (IAD) framework as a groundwork for developing a behavioral framework for social dilemmas. Section 3 lays out the basic structure and operation of the behavioral framework for social dilemmas that provides the foundations for the formal and empirical analyses conducted in Chapters 4 to 6. Section 4 presents a generic utility function as the foundation for modeling heterogeneous motivations. Section 5 reviews some of the extant models of heterogeneous motivations expressed in utility functions.

3.2 Review of Frameworks: MES and IAD

3.2.1 V. Smith's Microeconomic System Framework

Vernon Smith (1982) formalizes a framework that bridges the flourishing experimental studies and microeconomic theory. Smith himself does not consistently call it a framework. Overall, however, what he develops is clearly a framework, since, as he says, the main purpose of a microeconomic system is “to provide a taxonomy for laboratory experimentation which allows the methods, objectives and results of such experiments to be interpreted and perhaps extended” (923).

The MES framework is especially useful in studying how the characteristics of alternative institutions affect the aggregate social outcomes. By varying institutional rules while controlling for environmental variables, experimental research can draw conclusions on the relative performance of different institutions. In the MES framework, the microeconomic system is the unit of analysis.

A *microeconomic system* (S) consists of environment(e) and institutions(I):

$$S = (e, I). \quad (3.1)$$

Environment is considered to be initial circumstances that cannot be altered by the agents or the institutions.² *Environment*, e , consists of

- N agents $\{1, \dots, N\}$

²In a more general framework, control vs. focus variables need to be determined by research question and design. Any element in *Environment* can be theoretically and practically changed to examine its impact on behavior and aggregate social outcome.

$$\begin{array}{rcl}
 e & = & (u, T, \pi) \\
 \parallel & & \parallel \parallel \parallel \\
 (& & \\
 e^1 & = & u^1 T^1 \pi^1 \\
 e^2 & = & u^2 T^2 \pi^2 \\
 e^3 & = & u^3 T^3 \pi^3 \\
) & &
 \end{array}$$

Figure 3.1: Dual representation of an Environment

- $K + 1$ commodities (including resources) $\{0, 1, \dots, K\}$,
- certain characteristics of each agent i , $e^i = (u^i, T^i, \pi^i)$, where

u^i is agent i 's utility function,

T^i is agent i 's technology endowment, and

π^i is agent i 's commodity endowment vector.

An environment can be represented as a set of individuals with their characteristics

$$e = (e^1, \dots, e^N), \quad (3.2)$$

or as a set of population characteristics distributed among the individuals

$$e = (u, T, \pi). \quad (3.3)$$

In the first representation, an *Environment* is a set of individuals or simply a population. In the second representation, *Environment* consists of utility functions, technology, and endowment. Figure 3.1 shows the foundation for this dual representation with an *Environment* with three agents and three characteristics as an example.

Institution, I, defines “the rules of private property under which agents may com-

communicate and exchange or transform commodities for the purpose of modifying initial endowments in accordance with private tastes and knowledge” (924-5). An *Institution* specifies

- A language $M = (M^1, \dots, M^N)$ consisting of messages $m = (m^1, \dots, m^N)$, where m^i is an element of M^{i3}
- Allocation rules for each i $H = (h^1(m), \dots, h^N(m))$
- Cost imputation rules $C = (c^1(m), \dots, c^N(m))$, which could be included in H
- Adjustment process rules – starting rule, transition rule, stopping rule, all concerning exchange of messages.

In other words, an *Institution* defines the alternative choices (messages) each agent has at each stage of a transaction, the ways in which resources are allocated and costs are imputed to each agent as a result of the choices made by all the individuals, and the ways in which a transaction starts, proceeds, and ends. The *Institution* of a microeconomic system endows each individual with a set of alternative choices during each stage of transaction and a certain amount of resources as a result of transaction – that is why “defining property rights” is enough to capture the essence of the function of an *Institution*. Understood as such, an *Institution* is the collection of all these individual property right characteristics

$$I = (I^1, \dots, I^N), \quad (3.4)$$

³Message is analogous to strategy in game theory.

where each agent i 's property rights in communication and in exchange are defined by

$$\Gamma^i = (M^i, h^i(m), c^i(m), g^i(t_0, t, T)). \quad (3.5)$$

Or, alternatively, an *Institution* can be also defined as a set of rules that, in turn, defines each individual's property rights in the area of a transaction governed by each rule.

$$I = (M, h(m), c(m), g(t_0, t, T)) \quad (3.6)$$

where

$$M = (M^1, \dots, M^N) \quad (3.7)$$

⋮

etc.

The MES framework's high level of formality is particularly helpful in developing a behavioral framework for social dilemmas that can be used in conjunction with game theory. On the other hand, we will also see that the MES framework tends to assume away, or rather consciously detours, the problem of individuals' motivations. The existence of motivation(s) is a fact that cannot be disputed. The nature of motivations, however, is not directly observable.

This gives a conundrum in building a framework of which the working parts need to be empirical objects. Smith solves this problem by suggesting that a certain type of

preference can be induced through the devices contained in a series of precepts in the article.

The emphasis on preference inducement in the MES framework reflects the purposes Smith posited that were most important in the early days of experimental economics. The purposes of laboratory experiments, as Smith understood them, are

- to control the elements of a system $S = (e, I) = (u, T, \pi; M, h, c, g)$,
- to observe message responses of agents and the resulting outcomes, and
- to evaluate the performance of the system (930).

The primary purpose of experiments, thus, is the evaluation of relative performance of alternative institutions. This, of course, is one of the most important purposes of not only the laboratory but also the field research on any kind of human interactions. The problem is that this one purpose is given too much emphasis, and there follows a need and belief that all the elements in a system can be controlled in a well-devised experiment. Among the elements, it would certainly be possible to control the endowments(π), the messages (M), and the rules (h, c, g). But, is it also possible to control utility function(u) even in the laboratory? Smith proposes a set of precepts that he believes is sufficient to induce subjects' preferences. The four precepts are: *Nonsatiation*, *Saliency*, *Dominance*, and *Privacy* (931-35).

- **Nonsatiation:** Given a *costless* choice between two alternatives, identical (i.e., equivalent) except that the first yields more of a reward medium (for example, U.S. currency)

than the second, the first will always be chosen (i.e., preferred) over the second, by an *autonomous* individual.

- **Saliency:** Individuals are guaranteed the right to claim a reward that is increasing (decreasing) in good (bad) outcomes of an experiment; individual property rights in messages, and how messages are to be translated into outcomes, are defined by the institution of the experiment.
- **Dominance:** The reward structure dominates any subjective costs (or values) associated with participation in the activities of an experiment.
- **Privacy:** Each subject in an experiment is given information only on his/her own payoff alternatives.⁴

When a study accepts the precepts, it can be assumed that in every agent's utility function, utility is monotonically increasing in monetary reward and each agent chooses actions to maximize the expected monetary return. In sum, agents can be considered as rational self-interested actors.

In experiments examining the performance of market institutions, these precepts seem to work reasonably well. The precept of privacy deals with the possibility of "interpersonal utility considerations," the possibility that "individuals may not be autonomous own-reward maximizers" and, thus, their preferences diverge from what the experimenters intend to induce.

⁴Put in crude practical terms, these precepts require experimenters to pay enough monetary reward to the subjects, make the rewards directly related to their choices, keep the decision making costs quite low relative to the rewards, and keep the reward information private to each subject so that the subjects can consider only their own rewards.

It can be questioned whether or not the privacy condition is necessary and whether or not it can be achieved. First of all, inference on relative performance of institutions does not necessarily require that preferences be exactly controlled. Preferences can, and need to, be controlled in the sense that subjects are randomly assigned to different experimental conditions such that initial distributions of motivational tendencies across experimental treatments would not be biased. Once that control is achieved, different outcomes resulting from different institutions can be attributed to the institutions, whether or not the exact utility functions for all the individuals are controlled for and known to the experimenters.

Second, not providing full information regarding the reward structure of an experiment does not mean that the subjects cannot make a good conjecture about it. For example, in double continuous auction experiments with multiple rounds, most subjects quickly capture the overall demand and supply schedule after a few rounds. In that case, it is quite possible for them to infer how one's own choices affect the monetary rewards for others. The observed behavior in market experiments apparently indicates that preferences are induced successfully since the outcomes of such experiments are usually consistent with the theoretical predictions based on the universal self-interest assumption. But there are other possible interpretations for such observation (for example, Fehr and Schmidt, 1999). In the alternative explanations, it is the characteristics of the market institution that channel behavior to a certain direction regardless of the nature of individual preferences. The implication is that even when preference inducement fails, the institutional characteristics will induce certain behavior that gives an impression that the preference inducement has been successful.

The *Privacy* precept is particularly problematic in the experimental study of social dilemmas in which the experimenters have to make sure that each and every subject is aware of the interdependent reward structure. It is exactly how individuals consider and react to the interdependent nature of rewards allocation that researchers wish to know in the laboratory as well as in the field study of social dilemmas. In that sense, the *Privacy* precept is incompatible with the very purpose of the social dilemma research.

3.2.2 Institutional Analysis and Development Framework

The Institutional Analysis and Development (IAD) framework was developed by scholars associated with the Workshop in Political Theory and Policy Analysis at Indiana University synthesizing the teachings of classical political economy, public choice theory, new institutional economics, and non-cooperative game theory. It has been applied to formal and non-formal studies of macro and micro social phenomena. The IAD framework identifies a set of universal working parts and their generic relationships involved in intentional human interactions. It provides a common explanatory scheme for a wide range of action situations including markets, hierarchies, and social dilemmas (Ostrom, Gardner, and Walker, 1994:25-28). The basic conceptual unit of analysis in the IAD framework is an *action arena*, which includes an *action situation* component and an *actor* component. Figure 3.2 summarizes the basic components of an action arena in the IAD framework.

The specific characteristics of an action arena are generated and shaped by the *rules* that individuals use to order their relationships combined with the attributes of the *physical* and *cultural* world. Social outcomes result when actors, with preferences and capabilities, take actions within the constraints and opportunities posed by the rules and

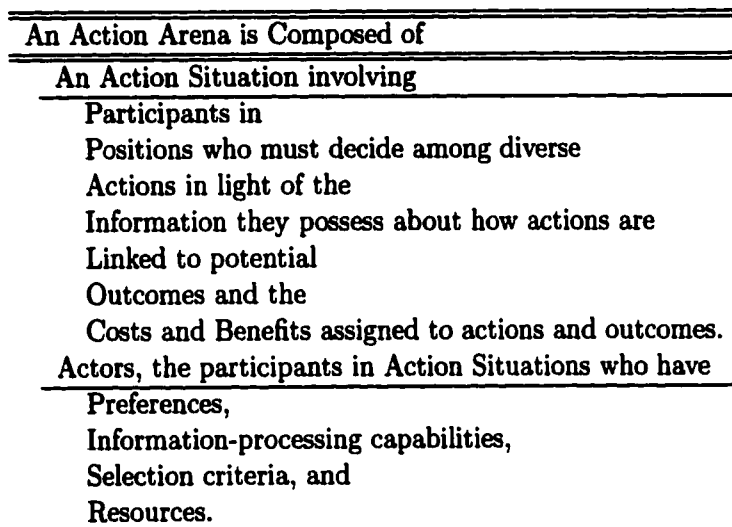


Figure 3.2: Components of Action Arenas. Source: Ostrom, Gardner, and Walker (1994: 29).

the physical and cultural world.

Though the working parts of a social space are categorized differently between the MES framework and the IAD framework, and the MES framework employs a formal expression for each working part while the IAD framework relies more on informal concepts, each individual working part in one framework easily finds its counterpart in the other. For example, *actor* in the IAD framework is conceptualized as *agent* in the MES framework. While the actor in the IAD framework is a higher-level category that encompasses multiple working parts of preference, information capabilities, selection criteria, and resources, the agent in the MES framework is an element of environment. However, it is not difficult to notice that both the frameworks attribute an individual in a social space of interaction as having preferences (utility functions), information-processing capabilities (technology), and resource (commodity) endowments.

Language (M) and adjustment processing rules (G) in the MES framework, defined

for each individual, specify each individual's *position* and available *actions* and *information* in the IAD framework. Allocation rules (*H*) and cost imputation rules (*C*) in the MES framework correspond to the elements of potential outcomes, transformation function, and payoffs in the IAD framework. While the seven elements of an action situation in the IAD framework are explicitly posited independent of the rules that define the elements, the corresponding elements in the MES framework are implicit in the definition of *Institution*.

Compared to the MES framework, the IAD framework is more flexible in dealing with the problem of unobservable preferences. The IAD framework does not require fixed assumptions (induced or speculative) on the nature of the preferences individuals might have. To the contrary, the IAD framework takes theories of preference and selection criteria as hypotheses and tests their validity based on the observed behavior of the agents. The behavioral framework of social dilemmas to be developed will follow the approach of the IAD framework in dealing with the problem of unobservable preferences.

3.3 A Behavioral Framework for Social Dilemmas and Its Operations

The striking commonality, in spite of some important differences, between the two frameworks reviewed in the previous section serves as the foundation for developing a behavioral framework for social dilemmas in this section. The essence is that the framework needs to help structure the social space in which human beings make decisions within a set of artificial and physical constraints. In terms of terminology, the framework will borrow extensively from the MES framework. This is because the formal and succinct nature of

the MES helps in developing a framework whose main purpose is to bridge the empirical world of social dilemmas to behavioral game theory. Substantively, however, the behavioral framework for social dilemmas can be considered as a formalized, simplified, and adjusted IAD framework for formal research on social dilemmas. As a limited framework, it will not incorporate the multiple levels of analysis in the IAD framework: constitutional, collective choice, and operational. It is also limited in that it assumes, if not specifies, individual utility function; the framework assumes that individuals do have preferences with certain characteristics. On the other hand, by simply replacing the utility function with other kinds of decision-making mechanisms, the framework can easily be made compatible with other theories.

3.3.1 A Behavioral Framework of Social Dilemmas

The social space in which interactions among individuals occur is called a system,

Δ . A system consists of *Environment*, e , and *Institution*, I :

$$\Delta = (e, I). \quad (3.8)$$

Environment, e , consists of

- N individuals $\{1, \dots, N\}$,
- K commodities (including resources), and
- certain characteristics of each individual i , $e^i = (u^i, T^i, \pi^i, \mu^i, q^i)$, where

u^i is individual i 's utility function,

T^i is individual i 's technology endowment,

π^i is individual i 's commodity (resource) endowment vector,

μ^i is individual i 's belief, and

q^i is individual i 's learning function that updates utility function,

technology, and belief.

An *Environment* can be defined in two ways. As a set of *environments* possessed and/or experienced by each individual

$$e = (e^1, \dots, e^N) \quad (3.9)$$

or, as a set of generic elements

$$e = (u, T, \pi, \mu, q). \quad (3.10)$$

Institution, I , specifies a set of actions $A = (A^1, \dots, A^N)$ where A^i is the set of actions available to agent i , a set of rules $R = (R_1 \dots R_L)$ that may include allocation rules R_a , communication rules R_c , information rules R_d , rules governing sanctioning R_s , rules governing monitoring R_m , etc., and a set of higher-order rules governing institutional change R' . Classification of rules is not attempted in the framework. It suffices to note that *Institution* specifies actions available to each individual – the rules of transaction and institutional change.

An *Institution* can be defined either as a sum of institutions defined on each of the N individuals

$$I = (I^1, \dots, I^N) \quad (3.11)$$

where

$$I^i = (A^i, R^i, R^i), \quad (3.12)$$

or as a combination of *institutions* governing separate substantive aspects of a transaction:

$$I = (A, R, R). \quad (3.13)$$

3.3.2 Operation of the Working Parts in the Framework

To present the operation (logical and temporal unfolding) of a system, a generic notation

$$X \quad \underline{Y} \quad Z \quad (3.14)$$

will be used. Expression (3.14) means “ Y maps X into Z ” or “ X is mapped into Z by Y .” Mapping implies either material transformation or logical consequence. For example, when an original system (X) is mapped into a resulting system (Z) through the function of an element (Y), the mapping is a material transformation. Or, when individuals make their choices (X), and a relevant set of rules (Y) in the system allocates resources among individuals and generates a new state of resource allocation (Z) based on the choice made by individuals, the mapping is the functioning of the rules in aggregating individual choices

into a scheme of resource allocation. In either case, Y , the medium, is considered as the cause of the material or logical correspondence between X and Z .

A system is activated when individuals choose actions. An individual i 's *behavior* is his/her selection and execution of an action, a^i , or a series of actions, $a^i = (a_1^i, \dots, a_t^i)$ over time, from the action set A^i . An individual's behavior is conditioned, in principle, by all of the elements in the system including his/her own characteristics defined in the previous subsection. In practice, which elements in the system are posited to condition an individual i 's behavior in what manner depends on the nature of theory used to animate the framework to explain observed functioning and consequences of a system. Since the explanatory power, or the empirical validity, of a theory is a crucial criterion in choosing which theory to use in conjunction with the framework, the relevant set of elements and the manner in which they condition individuals' behavior are empirical questions as well.

The fact that behavior is somehow conditioned is an essential requirement in scientific research on human social interactions. Otherwise, if behavior is purely random, it is meaningless to develop frameworks, theories, and models to explain it. Conditioned behavior implies the existence of a behavioral function β^i for each individual i , that maps the individual's own characteristics, e^i , into an action, a^i , conditional on environment, institution, and other individuals' behavior⁵:

⁵Compare (3.15) with Smith's behavioral function $\beta^i(e^i|I)$, which is non-strategic. It is non-strategic in the sense that it is not conditioned on others' behavior and characteristics. The point is not that behavior is always strategic in the sense of game theory, but that Smith's behavioral function tends to preclude strategic behavior. The generic behavioral function specified in this subsection, in turn, does not mean that all behaviors are strategic. It only allows the possibility. Parametric behavior can be modeled by minimizing the conditioning power of a certain subset of elements in the system.

$$(e^i | I, e, a^j) \xrightarrow{\beta^i} a^{i*} . \quad (3.15)$$

When viewed in aggregate, the behavioral function constitutes a process in which environment and institution are jointly mapped into a set of actions taken by individuals.

$$(I, e) \xrightarrow{\beta} a^* \quad (3.16)$$

where $\beta = (\beta^1, \dots, \beta^N)$ and $a^* = (a^{1*}, \dots, a^{N*})$.

A relevant set of rules in the rule set R (re)allocates resources among individuals based on the actions taken by agents, generating a new distribution of resource endowments among agents, π' . The set of actions actually taken by the individuals (a^*) is a consequence of the original system Δ .⁶ The resulting allocational state π' belongs to the new system Δ' in which the parameter values for the elements are different from those in the original system Δ . Therefore, it is now necessary to introduce a time dimension to the operation of the system and its elements. Outcome actions taken in the original system at time t , first create a new allocational state of affairs. And the new allocational state of affairs in turn creates and belongs to a new system at time $t + 1$. The relevant rules in the rule set at time t , R^t dictates the mapping:

$$a^{*t} \xrightarrow{R^t} \pi^{t+1} \quad (3.17)$$

⁶We could call it an allocation rule and have it subscripted, for example, R_a . But for the simplicity of presentation, further sub-classification of rules is not attempted in the notation. The analytical focus on utility function in this study does not require further classification.

where $\pi^{t+1} = (\pi^{1,t+1}, \dots, \pi^{N,t+1})$.

Changes in individual i 's preference, belief about the elements in the system, and technological and informational capabilities may occur when the individual i experiences and evaluates the consequence of his/her own action in the new state of allocation. Individuals will experience the allocational consequence. Whether or not, and on what level of concreteness, an individual will be informed about the intermediate factors that caused the new state of allocation – for example, actions actually taken by other individuals – depends on the relevant set of rules in the rule set R^t and his own reasoning capability defined in $T^{t,i}$. Each individual's learning function updates his utility function, technology, and beliefs based on a^* and π' . The learning function for an individual i dictates the process.⁷

$$(a^{*,t,i}, \pi^{t+1}) \quad \underline{q}, \quad (u^{t+1,i}, T^{t+1,i}, \mu^{t+1,i}) \quad (3.18)$$

In aggregate,

$$(a^{*,t}, \pi^t) \quad \underline{q}, \quad (u^{t+1}, T^{t+1}, \mu^{t+1}). \quad (3.19)$$

Therefore, it can be said that institutional rules regarding allocation and individuals' learning function jointly create a new environment based on individuals' behavior in previous time periods:

$$a^{*,t} \quad \underline{R^t, q} \quad e^{t+1}. \quad (3.20)$$

⁷It is possible that the learning function q itself is subject to change. However, to avoid complications, q is assumed to be constant over time. So, it does not carry temporal subscript.

The whole process is a transformation of the original system into a new one:

$$(\Delta : I, e)^t \xrightarrow{\beta^t} a^{*,t} \xrightarrow{R^t, q} (\Delta : I, e)^{t+1}. \quad (3.21)$$

When the media of the system transformation are suppressed, a holistic representation is possible in which a system as a whole is the ultimate cause of the new system:

$$(\Delta : I, e)^t \rightarrow (\Delta : I, e)^{t+1}. \quad (3.22)$$

3.3.3 Theories and Inference on Motivation

In a holistic sense, the initial system $(\Delta : I, e)^t$ is the cause of the immediately following system $(\Delta : I, e)^{t+1}$. In a micro perspective, differences in the resulting systems $(\Delta : I, e)^{t+1}$ are explained in terms of the differences in the original systems $(\Delta : I, e)^t$. When a subset of elements in (I, e) , call it x , is varied while all others are controlled, the differences in the resulting systems can be explained by the difference in the varied elements of the original systems:

$$(\Delta : I, e)_1^t \xrightarrow{x_1} (\Delta : I, e)_1^{t+1} \quad (3.23)$$

$$(\Delta : I, e)_2^t \xrightarrow{x_2} (\Delta : I, e)_2^{t+1}. \quad (3.24)$$

The two system transformations of (3.23) and (3.24) imply that all other elements in the two original systems were the same and the only difference was in the set of elements

x : $x = x_1$ in system 1 and $x = x_2$ in system 2. When the differences in x in the original systems do not make any difference in the resulting system, the element x is said to be irrelevant or inconsequential in explaining system transformation.

The mappings in the operation of a system have been presented in the most abstract form without specifying concrete mechanisms of the mappings. Theories provide the mechanisms. For example, non-cooperative game theory provides a specific case of the mapping of environmental and institutional characteristics of a system into individuals' outcome behavior. The assumptions of expected utility maximization and common knowledge in the standard non-cooperative game theory provide a specification of the mapping (3.15) that is repeated as (3.25) below:

$$(e^i | I, e, a^j) \xrightarrow{\beta^i} a^{i*} . \quad (3.25)$$

Whether or not a theory will be adopted to provide the concrete mapping mechanisms of a system depends, to a significant part, on the empirical validity of the theory in regard to the observable functioning of the system. To use a theory to operationalize the elements in a system, the elements need to be transformed into the elements of the implicit framework of the theory. Suppose we have a system $\Delta = (e, I) = (u, \pi, \mu, A, R)$ and are interested in explaining individuals' outcome behavior $a^* = (a^{1*}, \dots, a^{N*})$:

$$(\Delta : u, \pi, \mu, A, R) \longrightarrow a^* . \quad (3.26)$$

Theory "*explains*" why certain types of behavior, a^* , result from a system with certain

characteristics, $(\Delta : u, \pi, \mu, A, R)$. But to acquire the explanation, a correspondence between the empirical elements of the framework and the logical elements of the theory is necessary. For example, suppose we wish to provide an explanation based on a simple non-cooperative normal form game. The system Δ , then, needs to be transformed into a normal form game Γ , of which the working parts are the set of players $i \in I$, the pure strategy space S_i for each player i , and payoff functions f_i that give player i 's von Neumann-Morgenstern utility $u^i(s)$ for each strategy profile $s = (s_1, \dots, s_I)$. A game Γ is defined by I , S , and F :

$$\Gamma = (I, S, F) \quad (3.27)$$

where $I = (1, \dots, N)$, $S = (S_1, \dots, S_N)$, and $F = (f_1, \dots, f_N)$.

Therefore, an explanation of behavior using noncooperative game theory requires a transformation.⁸

$$(\Delta : u, \pi, \mu, A, R) \Rightarrow (\Gamma : I, S, F) \quad (3.28)$$

The transformation provides an explanation because noncooperative game theory has a well-defined mechanism by which initial conditions of a game generate equilibria:

$$\Gamma(I, S, F) \rightarrow s^* \quad (3.29)$$

In other words, (3.29), the way game theory solves a game, is an explanation of (3.25), the way the elements of a social dilemma generate actual behavior of individuals.

⁸The triple arrow (\Rightarrow) implies *correspondence*.

Unfortunately, the transformation is not automatic. Problems arise when one tries to map the five elements of a system into the three elements of a game. The action set, A , can be directly translated into the strategy set, S , of the game. Players are also evident. The main problem is how to construct payoff function, F . Empirical construction of a payoff function requires the knowledge of how actions taken by individuals assign utilities to each individual. Expression (3.30) is the way the mapping of actions to utility occurs in the framework. The set of actions taken by individuals (a) generates a new allocation of resources (π) among the individuals dictated by the relevant set of rules in the rule set R . Utility function, u , transforms a state of allocation to utility ($u(\pi)$):

$$a \xrightarrow{R} \pi \xrightarrow{u} u(\pi). \quad (3.30)$$

In game theory, this empirical process is abbreviated as

$$s \xrightarrow{F} u(s). \quad (3.31)$$

In the terminology of the framework, this is a direct mapping of actions to utilities. To construct an empirically valid payoff function of a game, attention needs to be paid to each step in (3.30). Action (a) is observable. The institution, R , is also observable. That allows us to map actions (a) to an allocation (π). To convert π into the payoffs in the normal form game, one needs to know utility functions of individuals, u , which are not directly observable. The essence of behavioral approach to motivation is the empirical inference on $u = (u^1, \dots, u^N)$, based on observable elements in the framework, Δ , and the observed

behavior of individuals, a^{*910} :

$$(\Delta, a^*) \implies u. \quad (3.33)$$

3.4 Generic Utility Function

3.4.1 Motivational Factors: A Generic Utility Function

Inference on u is in fact an inference on the utility functions of all the agents: (u^1, \dots, u^N) . The inference is based on the observable elements of the system, observed behavior of the agents, and the additional assumptions related to the application of game theoretic solution concepts: that individuals (agents, players) have utility functions that satisfy von Neumann-Morgenstern utility axioms and agents hold common and truthful knowledge regarding the structure of the system including the utility functions of others. Under most circumstances, identifying a unique $u = (u^1, \dots, u^N)$ would be an impossibility. In many cases, one can only narrow down the possibilities by eliminating some obviously wrong candidates. To retain scientific value, the inference has to be made on the generic utility function that has substantive meaning, applies to a wide variety of action situations,

⁹The double arrow (\implies) implies direction of inference.

¹⁰Game theoretic models with the self-interest assumption posits $u^i(\pi^i) = \pi^i$. With that assumption, a set of actions can be mapped into a set of von Neumann-Morgenstern utilities for individuals. Payoff function is defined. The common knowledge assumption provides the last requirement for the application of solution concepts in non-cooperative game theory:

$$\mu^1 = \mu^2 = \dots = \mu^N. \quad (3.32)$$

With the assumptions, a subset a^* can be selected from a . This is a mathematical/logical practice on purely a theoretical level and a prediction when applied to an observable action situation. Suppose the prediction fails. At a deeper level, one can consider alternative aggregation methods such as individual decision-making theory or evolutionary theory with their implication for construction of the system itself. Or, one might still wish to use the solution concepts of non-cooperative game theory as the primary mechanism of mapping the initial system to actions.

and is falsifiable.

Generic Utility Functions

While the payoff function, as an element of a normal form game, maps each strategy profile into the von Neumann-Morgenstern utility for each player, the utility function discussed here transforms each end state, usually π' , into each agent's von Neumann-Morgenstern utility. A generic utility function consists of a vector of arguments Π and a vector of parameters Θ :

$$u^i = u^i(\Pi; \Theta^i). \quad (3.34)$$

Inter-individual heterogeneity is captured by the individual specific values of the parameters, Θ^i . The game models based on the self-interest and risk neutrality assumptions utilize a specific operationalization of the generic utility function in which $\Pi = (\pi^i)$ and $\Theta^i = (1)$:

$$u^i = u^i(\pi^i) = \pi^i. \quad (3.35)$$

It is implied that individuals are selfish in the sense that they care only about their own material well-being and they are risk neutral. The self-interest assumption without the risk-neutrality assumption is a bit more general:

$$u^i = u^i(\pi^i), \quad \text{where } \frac{du^i}{d\pi^i} > 0. \quad (3.36)$$

It is assumed that individuals care only about their own material well-being and they

prefer a higher material payoff to a lesser one, regardless of how the other factors vary as a consequence.

3.4.2 Arguments in Utility Function: Factors Affecting Preference

Arguments in a utility function, and the ways the arguments and parameters combine, determine the mathematical representation of an individual's motivation. In addition, these two factors also determine whether the utility function supports a game; that is, whether game theory in the sense of von Neumann and Morgenstern (1953), Nash (1950), and Harsanyi (1967-68) can be used to model an action situation.

Others' Well-Being

In addition to one's own material wealth, what could be the additional arguments in an individual's utility function? The first and most natural candidate is $\pi^{j \neq i}$. Now individual i cares not only about the amount of the good available for her but also those for others. There are multiple possibilities concerning the relationship between π^i and $\pi^{j \neq i}$ in a utility function. Andreoni (1990), Cain (1998), Palfrey and Prisbrey (1997), Fehr and Schmidt (1999), and Bolton and Ockenfels (2000) give examples of utility functions in which $\pi^{j \neq i}$ is an argument.

Actions

Action, the selection and execution of a particular strategy, itself can be an argument in a utility function. Some people feel good when they do the things that they think they should do, even when doing so reduces material welfare. Some people feel bad – lose utility – when they do things that are inconsistent with the moral, social, or personal

principles they have, even when doing so increases their consumption of a commonly valued good. Crawford and Ostrom's (1995) δ parameter, and Palfrey and Prisbrey's (1997) warm-glow term, are the arguments in individual utility functions that directly reflect preferences over actions themselves, not the results of them. Placing action as an argument in utility functions raises two methodological questions. First, is this a violation of the instrumental rationality principle, often conceived as fundamental in rational choice and game theory? Second, is the action chosen an element of the outcome of a game? These two questions arise from the common understanding in game theory that preferences must be defined over the "outcomes" of a game. Or, outcomes of a game consist of all factors that affect players' expected utilities. Sen (1985) argues that action itself is often a crucial element in the state of affairs brought about by the action.¹¹ Though important, we can avoid these questions since there are ways to incorporate actions into the utility functions while meeting the mathematical requirements of a game.

Beliefs

Beliefs have been an essential element of game theory since Harsanyi's (1967-68) seminal article on games of incomplete information. But in standard game theory, beliefs' importance lies in the calculation of players' expected utilities. In Rabin's (1993) model of fairness, which will be introduced in the next section, belief is an argument in players' utility functions. Inclusion of beliefs into the utility function raises a question regarding the model's compatibility with the standard game theory.

¹¹"A state of affairs in which Brutus kills Caesar is not just one in which Caesar has expired. It is one in which the killing of Caesar by Brutus (and others) figures" (181).

Identity and History

Identity of co-players (Kollock, 1998) and history of relationship (Wagner, 1998) are also considered as arguments in utility functions. Kollock argues that the preferences over outcomes differ depending on the identity of the co-player(s) of a game. In Wagner's formal model of reciprocity, utility for a player varies depending upon past interaction in the game, holding material outcomes constant.

3.5 Non-selfish Utility Functions

The kinds of arguments, and the ways arguments and parameters are combined, define the substantive meaning of a utility function. Additional characteristics of a non-selfish utility function depend on scale sensitivity, symmetry, linearity, and other mathematical properties. Below, several non-selfish utility functions, proposed by behavioral game theorists and experimental social scientists, are introduced and their characteristics are briefly discussed. The concepts used to characterize these utility functions, such as altruism, fairness, equity, reciprocity, etc., themselves do not have clearly defined and universally accepted meanings. Thus, mathematical implementation of these concepts depends on the inventor's own interpretation of the concepts. To achieve notational coherence across different models, utility functions are modified without changing the inventor's original intentions. For simplicity, all utility functions are presented in the context of two-person action situations.¹²

¹²Two of the utility functions – the altruism function and the inequity aversion function – will be analyzed in detail in the next chapter.

3.5.1 Altruism

Altruism in its operationalized form often involves weights an individual gives to others' wealth (Jencks, 1990:53). The weights vary across individuals. Cain (1998) provides a model of the *Prisoner's Dilemma* in which players have different levels of altruism. In the model, an individual's utility function is specified as

$$u^i(\pi^i, \pi^j; \theta^i) = \pi^i + \theta^i \pi^j \quad (3.37)$$

where $0 \leq \theta^i \leq 1$. It is implied that $\Pi = (\pi^i, \pi^j)$ and $\Theta^i = (\theta^i)$. Notice that individual i 's utility monotonically increases, except when $\theta^i = 0$, not only in her own but also the other individual's material well-being. Therefore, θ^i can be interpreted as individual i 's rate of substitution between one's own and the other person's material payoffs.

3.5.2 Inequity Aversion

In Fehr and Schmidt's (1999) inequity aversion model, individual i 's preference over possible outcomes of a two-person game is represented by a utility function

$$u^i(\pi^i, \pi^j; \alpha^i, \beta^i) = \pi^i - \alpha^i \max(\pi^j - \pi^i, 0) - \beta^i \max(\pi^i - \pi^j, 0) \quad (3.38)$$

where it is assumed that $\beta^i \leq \alpha^i$ and $0 \leq \beta^i < 1$. It is implied that $\Pi = (\pi^i, \pi^j)$ and $\Theta^i = (\alpha^i, \beta^i)$. When both α^i and β^i equal zero, individual i 's preference is exactly the same as the one in the traditional models of self-interest. What the model of inequity aversion implies is that some, but not all, individuals prefer equal outcomes to unequal

ones, holding the level of one's own material payoffs constant. Outcomes of different games are explained in terms of the interaction between the distribution of types in a population and the characteristics of institutions that shape the interactions among individuals. This is an inequity aversion utility function in the sense that given an individual i 's material payoff of π^i , the other individual j 's material payoff π^j that maximizes i 's utility is $\pi^j = \pi^i$.

The preferred choice set to any given status quo for an individual with inequity aversion utility function is always a subset of that for the purely selfish individuals. When an increase in one's own wealth is accompanied by a substantial decrease in the other person's wealth, or when an increase in one's own wealth is accompanied by a far larger increase in the other person's wealth such that the level of inequity also increases more than a tolerable extent, the individual rejects the alternative even when it guarantees that he will be materially better off.

3.5.3 ERC (Equity, Reciprocity, Competition)

Bolton and Ockenfels (2000) argue that individuals are motivated by both the pecuniary payoff and the relative payoff standing. Their model of Equity, Reciprocity, and Competition (ERC) can be exemplified in two-person games by a utility function

$$u^i(\pi^i, \pi^j; \gamma^i) = \pi^i - \gamma^i \left(\frac{\pi^i}{\pi^i + \pi^j} - \frac{1}{2} \right) \quad (3.39)$$

where γ^i is the weight attached to the relative pecuniary payoff standing. The bigger γ^i is, the more concerned individual i is for equity. When γ^i is zero, individual i 's utility function is the same as the conventional self-interest utility function. Bolton and Ockenfels' model

is quite similar to that of Fehr and Schmidt. The key difference is that (1) Bolton and Ockenfels' utility function is nonlinear, which allows better application to dictator games but introduces additional complexities in calculation; and (2) it is based on the premise that as the size of the pecuniary payoff increases, individuals' preference becomes more selfish, – which is not the case in Fehr and Schmidt's model.

3.5.4 Fairness

Rabin's (1993) model is based on the observation that people are kind to those who help them, but want to hurt those who are unfair to them, even when doing so is not strictly in their material self-interest. A mathematical model that incorporates this persuasive observation, however, is rather complicated. In a two-person game with players 1 and 2, Player 1's subjective expected utility is a function of three factors:

$$u^1 = (a_1, b_2, c_1) \tag{3.40}$$

where a_1 is 1's action, b_2 is 1's belief about 2's action, and c_1 is 1's belief about 2's belief about 1's action (= 1's belief about b_1).

Let $\Pi(b_j) \equiv \{(\pi_i(a, b_j), \pi_j(b_j, a)) \mid a \in S_i\}$ denote the set of all feasible material payoffs player i can select given i 's belief b_j . Let $\pi_j^h(b_j)$ denote player j 's highest material payoff in $\Pi(b_j)$, and $\pi_j^l(b_j)$ the lowest material payoff among points that are Pareto-efficient in $\Pi(b_j)$. Equitable payoff $\pi_j^e(b_j) = [\pi_j^h(b_j) + \pi_j^l(b_j)]/2$ is a reference point against which to measure how generous player i is being to player j . Let $\pi_j^{\min}(b_j)$ denote the worst possible

payoff for player j in the set $\Pi(b_j)$. Player i 's kindness to player j is given by a formula

$$f_i(a_i, b_j) \equiv \frac{\pi_j(b_j, a_i) - \pi_j^e(b_j)}{\pi_j^h(b_j) - \pi_j^{\min}(b_j)}. \quad (3.41)$$

Player i 's belief about how kind player j is being to him is given by another formula

$$f_i^*(b_j, c_i) \equiv \frac{\pi_j(c_i, b_j) - \pi_i^e(c_j)}{\pi_i^h(c_i) - \pi_i^{\min}(c_i)}. \quad (3.42)$$

And the actual utility function of individual i is

$$U_i(a_i, b_j, c_i) \equiv \pi_i(a_i, b_j) + f_j^*(b_j, c_i)[1 + f_i(a_i, b_j)]. \quad (3.43)$$

Values of $f_i(\cdot)$ and $f_i^*(\cdot)$ lie in the interval $[-1, \frac{1}{2}]$. The bigger $f_i(\cdot)$ is, the kinder i is to j . Lower $f_i^*(\cdot)$ implies that player i believes that player j is treating him unfairly. Then, player i wishes to treat player j badly too, by choosing an action a_i , that makes $f_i(\cdot)$ low or negative.

The most interesting feature of Rabin's model is that it includes beliefs as arguments in players' utility function. Whether or not this is compatible with game theory is questionable.¹³ Falk and Fischbacher (1998) and Dufwenberg and Kirchsteiger (1998) provide models of reciprocity based on Rabin's theory, and extend its applicability to extended form games.

¹³Van Kolpin (1993) argues that one can apply conventional game theory to these games by including the choice of beliefs as additional parts of player' strategies.

3.5.5 Relationship Accounting

Wagner (1998) presents a dynamic model that is based on a social psychological process called “relationship accounting.” History, summarized as the relative material payoff standing between players in the past, is an argument of the utility function in the model. In two-person games played at time t , individual i 's utility function, in its simplest form¹⁴, is

$$u^{i,t} = (1 - \rho^i)\pi^{i,t} - \rho^i |\gamma^i G^{i,t-1} + (\pi^{i,t} - \pi^{j,t})| \quad (3.44)$$

where, G term summarizes the relative material payoff standing between the two players in the past, γ^i is the strength of memory (or the extent to which i thinks the past is important in making decisions for the current round), $\frac{\rho^i}{1-\rho^i}$ is i 's marginal rate of substitution between his material wealth and consideration for the norm, and $\pi^{i,t}$ is i 's material payoff at time t . If $G^{i,t-1} > 0$, i is the debtor to the creditor j . When an individual is in the debtor's position, giving more to the creditor is necessary to maximize the utility function. Type parameters γ^i and ρ^i decide the magnitude of self-disadvantageous inequality i has to endure to compensate the accumulated self-advantageous inequalities in the past.

¹⁴It is the simplest form in the sense that egalitarian norm is commonly accepted between the two players.

Chapter 4

Altruism or Equity?

4.1 Introduction

The models of nonselfishness introduced in Chapter 3 do not assert that specific kinds of motivations are universally held by individuals involved in a social dilemma. Instead, the models provide alternative *dimensions* of inter-individual heterogeneity.¹ For example, the altruism model does not assert that human beings are all altruistic. Rather, it proposes the dimension of altruism as a useful and empirically valid way of describing the different degrees of (non)selfishness that motivate individuals' behavior.

This chapter conducts theoretical and empirical investigations of two models of nonselfishness: altruism and inequity aversion. The two models are selected for in-depth analysis because

1. They are among the simplest alternative models to that of selfishness. Simplicity is a

¹Rabin's model is an exception. But it is not too difficult to modify the model into a model of motivational dimension.

valuable asset for any model. However, it is a principle of science that a complicated model with empirical validity be chosen over an elegant model that is empirically invalid. Therefore, the simplicity of the two models is not valued by itself but valued as a strategic merit in developing models of non-selfishness. The assumption of selfishness cannot be easily discarded in studying many types of action situations. It has been dominant too long in the study of social dilemmas. Therefore, this study builds and tests alternative models by incrementally relaxing the narrow selfishness assumption.

2. The two models are compatible with noncooperative game theory. Models specify dimensions for motivation. But motivation is only one factor that affects behavior. Theories specify other individual, institutional, and environmental factors and their relationships that result in the actual behavior of individuals. Noncooperative game theory is an alternative theory that can be used in explaining behavior in social dilemmas. Moreover, it is the most widely used theory. The priority given to noncooperative game theory in this study reflects the strategy of starting with the most simple and clear alternative and proceeding by means of incremental adjustment. When the best possible model based on noncooperative game theory reaches its limit, there will emerge a better ground on which to examine the validity of the fundamental assumptions of non-cooperative game theory and the possible directions to which disciplined adjustments can be made.

The altruism (equity) model proposes the dimension of altruism (equity) as the proper way of describing each individual's extent of (non)selfishness. In the following analyses, the meanings of the terms altruism and equity are no more or no less than what is shown

in the respective utility function. In other words, the concepts are used only to characterize the mathematical formulae of the utility functions(4.1) and (4.2) below. Both utility functions are presented in the context of two-person games in which π^i and π^j denote individual i 's and j 's material payoff, respectively.

- Altruism utility function

$$u^i(\pi^i, \pi^j; \theta^i) = \pi^i + \theta^i \pi^j \quad (4.1)$$

where $0 \leq \theta^i \leq 1$.

- Inequity aversion utility function

$$u^i(\pi^i, \pi^j; \alpha^i, \beta^i) = \pi^i - \alpha^i \max(\pi^j - \pi^i, 0) - \beta^i \max(\pi^i - \pi^j, 0) \quad (4.2)$$

where $\beta^i \leq \alpha^i$ and $0 \leq \beta^i < 1$.

The term altruism is often used to refer to any kind of nonselfishness because all kinds of nonselfish motivations involve a consideration for others' interest. However, altruism in this study is defined as a specific kind of nonselfishness often called "linear altruism" (Taylor, 1987; Cain, 1998, Dougherty and Cain, 1999). The principle of decision making reflected in the utility function can be viewed as a variant of the utilitarian principle (Harsanyi, 1955) in that it maximizes the weighted sum of individual welfare. The limiting case of the altruism utility function (4.1) toward the direction of selfishness (when $\theta^i = 0$) is the purely selfish utility function. The limiting case in the other direction (when $\theta^i = 1$)

is a fair utilitarian utility function in which a person weighs others' interest the same as one's own.²

The utility function (4.2) implies that individuals have different degrees of preference for equal allocation, of which the strength is denoted by the magnitude of the type parameters α^i and β^i . An individual with an aversion to inequality would sacrifice certain amounts of his/her own material payoff to increase or decrease others' material payoffs so as to achieve a more equitable allocation. The limiting case of equity-oriented preference toward selfishness is when both the parameters α^i and β^i take a value of 0 – in that case, the utility function degenerates into a purely selfish one. The other limiting cases are hard to define in the inequity aversion utility function because it contains two parameters and two restrictions. A case is when $\alpha^i = \beta^i$, indicating that an individual i 's aversion toward self-advantageous inequity is as strong as a self-disadvantageous inequity. Otherwise ($\alpha^i > \beta^i$), the function assumes that the latter is greater than the former. Another limiting case occurs when β^i approaches 1. The bigger β^i is, the stronger individual i 's aversion to inequitable allocation. A β^i almost as big as 1 implies that individual i is *almost* indifferent between keeping an amount of material payoff that exceeds others' and throwing it away to

²The altruism utility function in Cain (1998) has a form

$$U_i(X_\phi) = \phi_i \pi_i + \phi_{-i} \pi_j. \quad (4.3)$$

After dividing both sides of the equation by ϕ_i , the utility function is the same as the one used in this study. (Since utility is defined up to linear transformation, dividing the left side of the equation by any positive constant does not affect outcomes.)

Dougherty and Cain(1999) employ an altruism utility function in the form of

$$u_i = f_i(\pi_i + \theta_i \pi_j) \quad (4.4)$$

where f_i is any monotonically increasing function. Equation (4.4) adds, to the altruism function used in this study, the aspect of decreasing marginal utility of the weighted sum. In terms of ordinal ranking between two alternative allocations, (4.3) and (4.4) always result in the same preference insofar as the type parameters take the same value.

achieve strict equality.

Parameter range restrictions in each of the two models imply implausibility, not impossibility. For example, in the altruism function, cases with θ^i greater than 1 or less than 0 is imaginable. The former implies that individual i values others' well-being more than his own. The latter implies that any increase in others' well-being decreases one's own utility. Both are possible. But they are considered insignificant in explaining motivations and behavior in social dilemmas.

Before conducting a more formal analysis of the two utility functions, it helps to compare these two and other motivational models with a simple example. Given a constant level of one's own well-being, an altruistic individual's utility increases in others' well being. Suppose Mr. Smith obtains \$100 as a result of an interaction with another individual. If Mr. Smith was a person with some degree of aversion to inequity, the amount of money for the other person Mr. Smith preferred most is also \$100. Any more money going to the other person decreases his utility. If he was a person with a very strong aversion to inequity, he would prefer an equitable, but Pareto-inferior, allocation of \$50 to each, to an inequitable, but Pareto-superior, allocation of \$60 to himself and \$100 to the other. However, if he was an altruistic individual, as operationalized here, he would be willing to sacrifice some amount of his money if that substantially increases the other person's share. For example, a strong altruistic individual would prefer an allocation in which his share is \$40 and the other person's share is \$100 to an alternative allocation in which each receives \$50.

Figures 4.1 to 4.3 show, respectively, the indifference mappings of an individual with purely selfish motivation (Figure 4.1), with moderate altruism (Figure 4.2), and with

moderate aversion to inequity (Figure 4.3) on an allocation space between oneself and another person. In each of the figures, the x -axis represents the individual's own material payoff and the y -axis the other person's. In Figure 4.2, $\theta^i = 0.5$. In Figure 4.3, $\alpha = 0.5$ and $\beta = 0.3$.

A selfish individual, whose indifference mapping is shown in Figure 4.1, prefers an allocation in which his payoff is larger than another allocation with a lower payoff, regardless of what payoff the other person obtains. For example, he prefers an allocation of \$100 to himself and \$0 to the other person to another allocation in which each gets \$99. On the other hand, an altruistic individual, whose indifference mapping is shown in Figure 4.2, would reject some alternatives to the status quo in which his share increases if that increase is accompanied by too much decrease in the other person's share. The exact trade-off points for an altruistic individual is determined by his altruism parameter θ^i . Notice that any Pareto-improving reallocation is approved by an altruistic individual. On the other hand, an individual with some degree of aversion to inequity may not approve of a Pareto-superior reallocation if it increases the difference in the relative standing too much. For example, in Figure 4.3, an allocation point (2,5) falls in the left side of an indifference line that passes the allocation point (1,1), indicating that (1,1) is preferred to (2,5).

Since both the altruism model and the inequity aversion model incorporate heterogeneity, testing the relative performance of the models is not straightforward. It is not simply a question of "are individuals altruistic or averse to inequity?" The correct question is "Which is the better model of inter-individual heterogeneity?" Another important question is whether the relative performances of the models are specific to social dilemmas

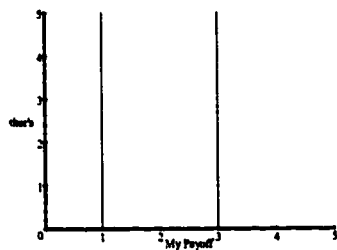


Figure 4.1: Indifference Mapping of Pure Selfishness

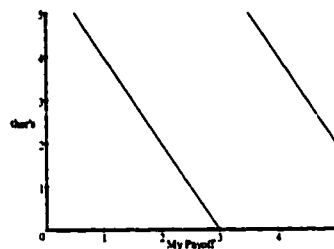


Figure 4.2: Indifference Mapping of Linear Altruism

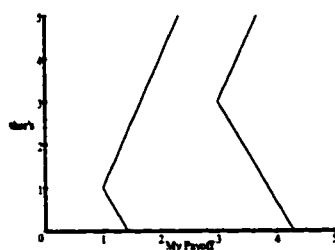


Figure 4.3: Indifference Mapping of Inequity Aversion

		<i>Individual 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Individual 1</i>	<i>Cooperation</i>	<i>R, R</i>	<i>S, T</i>
	<i>Defection</i>	<i>P, P</i>	<i>T, S</i>

* *T, R, P*, and *S* represent material payoffs
 ** $T > R > P > S$

Figure 4.4: 2×2 Social Dilemma

or generalizable to a wider scope of action situations. This would require a broader and more systematic empirical test of the models. In this study, the focus is on how the models perform in social dilemmas in which one person's gain can be the other's loss, while at the same time there exists a way for both to be better off by not blindly pursuing one's own self-interest. Therefore, any conclusions drawn from the empirical tests of this study do not extend to non-dilemma action situations.

Sections 2 to 4 derive theoretical implications of the two models regarding types of preference ordering (Section 2), equilibria (Section 3), and behavior in four information sets in the simultaneous and sequential 2×2 social dilemma games (Section 4). Section 5 tests the implications drawing on a set of experimental, two-person social dilemma games.

4.2 Preference-Ordering Types

Models restrict possible states of the world. A model that does not restrict at all may not be falsified. However, in that case, its value as a model does not exist since it adds no understanding of what is possible and what is not in the empirical world. Therefore, we will see how each of the two models restricts the possible state of affairs in meaningful ways and how their restrictions fit with the empirical evidence available.

Figure 4.4 recapitulates the two-person, binary choice social dilemma action situation, heretofore called 2×2 social dilemma. The concepts of *Fear*, *Greed*, and *Cooperators' Gain* (Rapoport and Chammah, 1965; Ahn et al., 2001) greatly simplify the discussion of the material payoff structure of a 2×2 social dilemma. When the other player cooperates, a player can increase his material payoff by defecting, by $T - R$, which is called *Greed*. He can also increase his material payoff by defecting when the other player defects, by $P - S$, called *Fear*. As a group, however, players can increase their material payoff by $R - P$, called *Cooperators' Gain*, by moving from mutual defection to mutual cooperation. When these three quantities are normalized by the range of possible payoffs ($T - S$), they are called *normalized Greed* (G_n), *normalized Fear* (F_n), and *normalized Cooperators' Gain* (C_n). The relationship among these three quantities characterizes a 2×2 social dilemma quite well, because the sum of these three measurements is always 1:

$$G_n(\text{Normalized Greed}) = \frac{T - R}{T - S} \quad (4.5)$$

$$F_n(\text{Normalized Fear}) = \frac{P - S}{T - S} \quad (4.6)$$

$$C_n(\text{Normalized Cooperators' Gain}) = \frac{R - P}{T - S} \quad (4.7)$$

$$G_n + C_n + F_n = 1. \quad (4.8)$$

Figure 4.4 itself does not represent a game since the payoff entries are material payoffs and not von Neumann-Morgenstern utilities. To convert the social dilemma action situation shown in Figure 4.4 to a standard game, the utility functions of individuals need

		<i>Player 2</i>	
		<i>C</i>	<i>D</i>
<i>Player 1</i>	<i>C</i>	$u^1(R, R), u^2(R, R)$	$u^1(S, T), u^2(S, T)$
	<i>D</i>	$u^1(T, S), u^2(T, S)$	$u^1(P, P), u^2(P, P)$

Figure 4.5: Normal Form Game Representation of a 2×2 Social Dilemma

to be specified. Figure 4.5 is the social dilemma *game*.³

In a purely selfish utility function

$$u^i = u^i(\pi^i) = \pi^i, \quad (4.9)$$

where π^i is individual i 's material payoff, only one preference ordering shown in (4.10) is possible.

$$u^i(T, S) > u^i(R, R) > u^i(P, P) > u^i(S, T) \quad (4.10)$$

Our focus in this section is what other types of preference do the models of altruism and inequity aversion allow.

4.2.1 Altruism

First, as is the case in both the models of selfishness and inequity aversion, altruistic individuals operationalized by the utility function of (4.1), prefer mutual *Cooperation* to mutual *Defection* regardless of the value of their type parameter θ^i :

$$u^i(R, R) > u^i(P, P). \quad (4.11)$$

³Strictly speaking, Figure 4.5 shows a game with complete information, which is different from the game with incomplete information to be analyzed in this chapter.

Proof.

$$u^i(R, R) > u^i(P, P)$$

$$R + \theta^i R > P + \theta^i P$$

$$(R - P) + \theta^i(R - P) > 0$$

$$(R - P)(1 + \theta^i) > 0 \quad (4.12)$$

Inequality (4.12) is always true because both $(R - P)$ and $(1 + \theta^i)$ are greater than 0. ■

Under what conditions would an individual i prefer to cooperate when the other player cooperates? The range of type parameter θ^i that satisfies this partial preference ordering can be calculated as follows:

$$u^i(R, R) > u^i(T, S) \quad (4.13)$$

$$R + \theta^i R > T + \theta^i S$$

$$\theta^i(R - S) > T - R$$

$$\theta^i > \frac{T - R}{R - S} \quad (4.14)$$

Rewriting inequality (4.14) using the previously introduced notations of normalized game parameters simplifies the expression. Inequality (4.14) can be rewritten as

$$\theta^i > \frac{T - R}{T - S - (T - R)} \quad (4.15)$$

Divide both the numerator and denominator of the right side of the inequality by the range of material payoff parameters, $(T - S)$:

$$\theta^i > \frac{G_n}{1 - G_n}. \quad (4.16)$$

If θ^i , individual i 's altruism parameter, is greater than $\frac{G_n}{1 - G_n}$, she prefers to cooperate when the other person also cooperates. Accordingly, if

$$\theta^i < \frac{G_n}{1 - G_n} \quad (4.17)$$

individual i prefers to defect when the other player cooperates. If

$$\theta^i = \frac{G_n}{1 - G_n}, \quad (4.18)$$

she is indifferent between *Cooperation* and *Defection* when the other player cooperates.

Is it possible, against our intuition, that an individual i prefers to cooperate even when the other player defects? If it is ever possible, a condition should be satisfied:

$$u^i(S, T) > u^i(P, P) \quad (4.19)$$

$$S + \theta^i T > P + \theta^i P$$

$$\theta^i(T - P) > P - S$$

$$\theta^i > \frac{P - S}{T - P}. \quad (4.20)$$

Inequality (4.20) can be simplified using the normalized payoff parameters. Rewrite (4.20) as

$$\theta^i > \frac{P - S}{T - S - (P - S)} \quad (4.21)$$

and divide both the numerator and denominator of the right side of the inequality by the range of material payoff parameters $(T - S)$. Then the condition can be expressed as

$$\theta^i > \frac{F_n}{1 - F_n}. \quad (4.22)$$

Therefore, if θ^i , individual i 's altruism parameter, is bigger than $\frac{F_n}{1 - F_n}$, he prefers to cooperate even when the other person defects. Accordingly, if

$$\theta^i < \frac{F_n}{1 - F_n}, \quad (4.23)$$

individual i prefers to defect when the other player defects. If

$$\theta^i = \frac{F_n}{1 - F_n}, \quad (4.24)$$

he is indifferent between *Defection* and *Cooperation* when the other player defects.

Possible preference orderings and their supporting parameter conditions can be derived based on the (in)equalities (4.16) to (4.18) and (4.22) to (4.24). There are four major behavioral types that are implied by the model of linear altruism: unconditional defection, unconditional cooperation, and two types of conditional cooperation. Conditions 5 to 8 provide supporting parameter conditions for these behavioral types.

Condition 5 Unconditional Defection: $\theta^i < \min[\frac{F_n}{1-F_n}, \frac{G_n}{1-G_n}]$

Individual i always defects regardless of the other player's choice: This implies that $\theta^i < \frac{G_n}{1-G_n}$ (4.17) and $\theta^i < \frac{F_n}{1-F_n}$ (4.23). In other words, $\theta^i < \min[\frac{F_n}{1-F_n}, \frac{G_n}{1-G_n}]$. Consistent with our intuition, when other players' material well-being does not figure much in one's utility function, the individual is less likely to cooperate. The backside of this condition is a relatively large temptation for defection expressed by the large normalized material payoff parameters F_n and G_n . The larger the two normalized parameters, the larger $\min[\frac{F_n}{1-F_n}, \frac{G_n}{1-G_n}]$ and it becomes less likely that an individual's consideration for other's well-being is big enough to overcome the temptation of defection present in the material payoff structure of a social dilemma.

Condition 6 Unconditional Cooperation: $\theta^i > \max[\frac{F_n}{1-F_n}, \frac{G_n}{1-G_n}]$

Individual i always prefers to cooperate regardless of the other player's choice. This requires $\theta^i > \frac{G_n}{1-G_n}$ (4.16) and $\theta^i > \frac{F_n}{1-F_n}$ (4.22). In other words, if $\theta^i > \max[\frac{F_n}{1-F_n}, \frac{G_n}{1-G_n}]$, *Cooperation* is individual i 's dominant strategy. In part, this result is consistent with the simple intuition that the more one cares about others' interest, the more likely he is to cooperate. The flip side of this condition is that the two kinds of temptations, the normalized *Fear* (F_n) and *Greed* (G_n), need to be small enough so that individual i 's altruistic orientation can overcome them. Again, intuitively, the smaller the material temptations of a social dilemma, the more likely an individual will cooperate.

Condition 7 Type 1 Conditional Cooperation: $\frac{G_n}{1-G_n} < \theta^i < \frac{F_n}{1-F_n}$

Individual i prefers to cooperate when the other player cooperates, but prefers

to defect when the other player defects. This requires $\theta^i > \frac{G_n}{1-G_n}$ (4.16) and $\theta^i < \frac{F_n}{1-F_n}$ (4.23). In other words, $\frac{G_n}{1-G_n} < \theta^i < \frac{F_n}{1-F_n}$. This type of preference can be understood as a conditional cooperation based on the reciprocity principle. A necessary condition for the existence of the preference type is

$$\begin{aligned} \frac{G_n}{1-G_n} &< \frac{F_n}{1-F_n} \\ G_n &< F_n. \end{aligned} \tag{4.25}$$

Inequality (4.25) as a necessary condition implies that in a social dilemma in which the structure of material payoffs is such that $F_n > G_n$, players can never have this type of preference.

Condition 8 Type 2 Conditional Cooperation: $\frac{F_n}{1-F_n} < \theta^i < \frac{G_n}{1-G_n}$

Individual i prefers to cooperate when the other player defects, but prefers to defect when the other player cooperates. This requires $\theta^i > \frac{F_n}{1-F_n}$ (4.22) and $\theta^i < \frac{G_n}{1-G_n}$ (4.17). In other words, $\frac{F_n}{1-F_n} < \theta^i < \frac{G_n}{1-G_n}$. The existence of this type of preference requires a necessary condition

$$\begin{aligned} \frac{F_n}{1-F_n} &< \frac{G_n}{1-G_n} \\ F_n &< G_n. \end{aligned} \tag{4.26}$$

Again, this implies that in certain kinds of social dilemmas with relatively large normalized

Type	Preference Ordering	Interpretation
<i>PD</i>	$u^i(T, S) > u^i(R, R) > u^i(P, P) > u^i(S, T)$	Always Defect
<i>Assurance</i>	$u^i(R, R) > u^i(T, S) > u^i(P, P) > u^i(S, T)$	Reciprocity
<i>Chicken</i>	$u^i(T, S) > u^i(R, R) > u^i(S, T) > u^i(P, P)$?
<i>Angel</i>	$u^i(R, R) > u^i(T, S) > u^i(S, T) > u^i(P, P)$	Always Cooperate

Table 4.1: Preference Types and Interpretation: Altruism Model

Type	Parameter Condition	Necessary Condition
<i>PD</i>	$\theta^i < \min\left[\frac{F_n}{1-F_n}, \frac{G_n}{1-G_n}\right]$	None
<i>Assurance</i>	$\frac{G_n}{1-G_n} < \theta^i < \frac{F_n}{1-F_n}$	$F_n > G_n$
<i>Chicken</i>	$\frac{F_n}{1-F_n} < \theta^i < \frac{G_n}{1-G_n}$	$F_n < G_n$
<i>Angel</i>	$\theta^i > \max\left[\frac{F_n}{1-F_n}, \frac{G_n}{1-G_n}\right]$	None

Table 4.2: Conditions for Preference Types: Altruism Model

Greed (G_n), players can never have this type of preference. Heretofore the four preference types will be called *PD* (*Prisoner's Dilemma*) *Angel*, *Assurance*, and *Chicken* preferences.⁴ Tables 4.1 and 4.2 summarize possible preference types and supporting conditions in the model of altruism.

4.2.2 Inequity Aversion

Both the models presented in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) can be regarded as models of inequity aversion. For a given 2×2 social dilemma game, the differences between the two models can be ignored.⁵ To simplify analysis, this

⁴Naming of Type 2 Conditional Cooperation as *Chicken* preference does not imply any sense of cowardice. It simply reflects that it is the standard preference type of players in the *Chicken* game.

⁵The main difference between the two models in the context of a given 2×2 social dilemma is that while in Fehr and Schmidt's model only the normalized material payoff parameters matter, in one interpretation of Bolton and Ockenfels model, the absolute material payoff parameters also matter. Another noticeable difference is the linearity (Fehr and Schmidt) versus nonlinearity (Bolton and Ockenfels). But this can be

section uses Fehr and Schmidt's model shown in (4.2) to represent the general family of inequity aversion utility function. In the model it is always the case that

$$u^i(R, R) > u^i(P, P) > u^i(S, T). \quad (4.27)$$

Note that this condition is always true in the model of pure selfishness, but not in the altruism model. Substantively, (4.27) means that every individual prefers mutual *Cooperation* to mutual *Defection* and mutual *Defection* to an outcome in which he/she is the lone cooperator. So the question is where $u^i(T, S)$ falls in.

Under what parameter conditions would an individual i prefer to defect even when the other player cooperates? For that, the following condition needs to be satisfied:

$$\begin{aligned} u^i(T, S) &> u^i(R, R) && (4.28) \\ T - \beta^i(T - S) &> R \\ \beta^i(T - S) &< T - R \\ \beta^i &< \frac{T - R}{T - S} \\ \beta^i &< G_n. && (4.29) \end{aligned}$$

Substantively, when β^i , the weight attached to utility loss due to self-advantageous inequity, is smaller than the normalized greed in a 2×2 social dilemma game, *Defection* is the dominant strategy of player i . Since the type parameter β^i has a lower bound of 0, we can

regarded as non-significant technical matter.

finalize the condition as:

Condition 9 *Unconditional Defection*: $0 \leq \beta^i < G_n$.

When Condition 9 is satisfied, a player has the dominant strategy of *Defection* and his preference is called a PD type. When Condition 9 is not met, two preference-ordering possibilities exist. A player with either of the two possible orderings can be reasonably called *Assurance* preference type in the sense that they prefer to cooperate when the other player cooperates, but prefer to defect when the other player defects. Since the type parameter β^i is bounded below 1, the condition can be finalized as:

Condition 10 *Conditional Cooperation*: $G_n \leq \beta^i < 1$.

Substantively, when the weight attached to utility loss due to self-advantageous inequity is greater than the material payoff parameter normalized gain (G_n), a player prefers to cooperate when the other player cooperates, but prefers to defect when the other player defects. *Assurance* preference type can be divided into two subtypes depending on whether the outcome in which one defects and the other cooperates is placed second or third in the ordering of the four outcomes. The outcome (T, S) is placed second if

Condition 11 *Type 1 Conditional Cooperation*: $G_n \leq \beta^i < G_n + C_n$.

The corresponding preference ordering is

$$u^i(R, R) > u^i(T, S) > u^i(P, P) > u^i(S, T). \quad (4.30)$$

Outcome (T, S) is placed in the third if

Condition 12 *Type 2 Conditional Cooperation*: $G_n + C_n \leq \beta^i < 1$.

The corresponding preference ordering is

$$u^i(R, R) > u^i(P, P) > u^i(T, S) > u^i(S, T). \quad (4.31)$$

Notice that compared to the model of altruism, the inequity aversion model results in a smaller number of preference-ordering possibilities. Specifically, it eliminates the possibility of *Angel* and *Chicken* types found in the altruism model. In a broader sense, assuming only strict orderings, the inequity aversion model allows only two preference types: the *Assurance* type (who is a conditional reciprocator) and the *PD* type (who is an unconditional defector).

Also notice that, contrary to the model of altruism, the inequity aversion model does not require any necessary condition, expressed in terms of certain relationships among the normalized material payoff parameters, for the existence of certain preference types. Substantively, this means that any of the three preference types is possible regardless of the relative magnitudes of the three material payoff parameters G_n , F_n , and C_n .

The conditions for preference types are expressed only in terms of β^i , the weight in one's utility function attached to the utility loss due to self-disadvantageous inequity. Parameter α^i plays a role in equilibrium analysis, but is not a determining factor of an individual's preference-ordering type for a 2×2 social dilemma game. Tables 4.3 and 4.4 summarize the possible preference-ordering types and supporting parameter conditions derived from the model of inequity aversion.

Type	Preference Ordering	Interpretation
PD	$u^i(T, S) > u^i(R, R) > u^i(P, P) > u^i(S, T)$	Always Defect
Assurance I	$u^i(R, R) > u^i(T, S) > u^i(P, P) > u^i(S, T)$	Reciprocity
Assurance II	$u^i(R, R) > u^i(P, P) > u^i(T, S) > u^i(S, T)$	Reciprocity

Table 4.3: Preference Types and Interpretation: Inequity Aversion Model

Type	Parameter Condition	Necessary Condition
PD	$0 \leq \beta^i < G_n$	None
Assurance I	$G_n \leq \beta^i < G_n + C_n$	None
Assurance II	$G_n + C_n \leq \beta^i < 1$	None

Table 4.4: Conditions for Preference Types: Inequity Aversion Model

4.3 Equilibria

This section analyzes the equilibria of the 2×2 social dilemma game. To apply the solution concepts of non-cooperative game theory, some assumptions are necessary. They are:

1. There are two players, 1 and 2, randomly drawn from a population denoted Θ , which can be characterized by a commonly known cumulative distribution function of the type parameters, $F(\theta^i)$.⁶ θ^i is individual i 's private information. However, $F(\theta^i)$ is a common knowledge; both players know $F(\theta^i)$, each knows that the other knows, each knows that he knows, etc.
2. Material payoffs are known with certainty and denoted by $\pi = (T, R, P, S)$.
3. The game is played once.

⁶The assumption that two players are randomly drawn from a population to play the 2×2 social dilemma game corresponds nicely to the actual experimental procedure used to generate the data.

In the inequity aversion model, θ^i is a vector with two elements α^i, β^i . In the altruism model, θ^i is a singular. It is an empirical question whether or not players have common knowledge about the distribution of types, or whether their beliefs converge to a common knowledge with learning in a repeated game setting. However, in this section, we assume that players do have common knowledge regarding the distribution of types within the population. This assumption is inevitable insofar as using the solution concepts of non-cooperative game theory. By deriving testable hypotheses from the equilibria analyses, we will have a chance to see what kind of anomalies this assumption generates, how significant they are, and how to modify them in a disciplined and empirically valid manner. Since the players do not know the exact payoff function of the other players, but only the probability distribution of types within the population, the game is of incomplete information.

Player i 's strategy depends on four factors:

1. His/her own type parameter: θ^i
2. The structure of material payoffs: $\pi = (T, R, P, S)$
3. The distribution of types within the population from which the players are drawn:
 $F(\theta^i)$
4. The sequence of play: whether the game is played simultaneously or sequentially. In a sequential game, it matters whether a player is the first or the second mover.

A player's own type and the structure of material payoffs jointly determine his preference ordering over the four possible outcomes of a 2×2 social dilemma game. The structure of the material payoffs and the distribution of types within the population jointly

determine a player's belief about the other player's strategy in an equilibrium. Finally, a player's preference ordering over the outcomes and his belief about the other player's likely strategy jointly determine his own strategy.

A Bayesian equilibrium of a 2×2 social dilemma game is a pair of strategies $(s_1^*(\theta^1), s_2^*(\theta^2))$ such that for each player i and every possible value of θ^i , strategy $s_i^*(\theta^i)$ maximizes expected utility, $E(u^i(s_i, s_j^*(\theta^j)))$. That is, in equilibrium, each player's strategy is the best response – in the sense of expected utility maximization – given his own type (θ^i) and the other player's strategy. Let us denote C and D as the strategies of *Cooperation* and *Defection*, respectively.

4.3.1 Inequity Aversion Model

Equilibria of Simultaneous Game

We first examine equilibria of the simultaneous game.

Proposition 13 $[s_i(\theta^i) = D \text{ for } i = 1 \text{ and } 2]$ is an equilibrium regardless of the structure of the material payoffs and the distribution of types within the population Θ . That is, there always exists an equilibrium in which both players Defect regardless of their types.

Excluding weak ordering, a player can have either *Assurance* preference ($u^i(R, R) > u^i(T, S)$) or *PD* preference ($u^i(T, S) > u^i(R, R)$). (D, D) is an equilibrium in all possible combinations of preference types of the two players.

Proposition 14 If there exists a subset Θ^c of the population Θ , such that

$$\beta^i \geq G_n \quad \text{and}$$

$$\frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i} \leq \Pr(i \in \Theta^c | i \in \Theta) \quad \text{for all } i \in \Theta^c,$$

$[s_i(i \in \Theta^c) = C, s_i(i \notin \Theta^c) = D]$ is an equilibrium.

That is, if there exists a subset Θ^c of the population Θ , in which everyone has an *Assurance* preference over the four outcomes and has reservation probability (μ^i) lower than the probability of a player in the population belonging to the subset (μ^*), an equilibrium exists in which all the types (defined in terms of θ^i) that belong to the subset Θ^c cooperate and the others defect. Reservation probability $\frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i}$, denoted μ^i heretofore, is the lower bound of the probability that the other player will cooperate with which a player with *Assurance* preference type cooperates in a simultaneous social dilemma game. The condition in Proposition 14 is repeated as

Condition 15 *Reservation Probability (μ^i) and Proportion of Cooperators (μ^*): Reservation Probability (μ^i) for each of the individuals in the Cooperative subpopulation should be smaller than the proportion of cooperators in the whole population (μ^*):*

$$\frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i} \leq \Pr(i \in \Theta^c | i \in \Theta) \quad (4.32)$$

Or simply

$$\mu^i \leq \mu^* \quad (4.33)$$

for all $i \in \Theta$.

Proof. The key in proving Proposition 14 is that it is possible, under certain conditions, that a subset of *Assurance* preference type players cooperate in spite of the existence of unconditional defectors, the *PD* type players. In that sense, the existence of

a cooperative equilibrium depends on a self-fulfilling subpopulation. Assume that there exists such a subpopulation, Θ^c , and see what conditions should be met. In the cooperative equilibrium, there is a μ^* ($= \Pr(i \in \Theta^c | i \in \Theta)$) probability that one's partner in the game will cooperate. And, a player cooperates if and only if the expected utility from *Cooperation* $Eu^i(C)$ is greater than or equal to the expected utility from *Defection*, $Eu^i(D)$. Or,

$$\begin{aligned}
 Eu^i(C) &\geq Eu^i(D) \\
 \mu^*[u^i(R, R)] + (1 - \mu^*)[u^i(S, T)] &\geq \mu^*[u^i(T, S)] + (1 - \mu^*)[u^i(P, P)] \\
 \mu^*(R) + (1 - \mu^*)(S - \alpha^i(T - S)) &\geq \mu^*(T - \beta^i(T - S)) + (1 - \mu^*)P \\
 (T - S)\mu^*(\alpha^i + \beta^i) + \mu^*[(P - S) - (T - R)] &\geq (P - S) + \alpha^i(T - S). \quad (4.34)
 \end{aligned}$$

To simplify the inequality using the normalized material payoff parameters, divide both sides of (4.34) by the normalization scale $(T - S)$.

$$\begin{aligned}
 \mu^*(\alpha^i + \beta^i) - \mu^*G_n + \mu^*F_n &\geq F_n + \alpha^i \\
 \mu^*(\alpha^i + \beta^i - G_n + F_n) &\geq F_n + \alpha^i \\
 \mu^* &\geq \frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i} \\
 \mu^i &\leq \mu^*. \quad (4.35)
 \end{aligned}$$

(4.35) is a repetition of condition (4.33). ■

Notice that Proposition 10 does not imply that whenever there exists a subpopulation of *Assurance*-type players there will always exist a cooperative equilibrium. It is only

when there exists a subpopulation in which all the individuals have small enough reservation probability (meaning higher willingness to cooperate), that there exists a cooperative equilibrium in which a subset of *Assurance* preference type players cooperate. Reservation probability of an individual i is determined jointly by his own type parameters, $\alpha^i + \beta^i$, and the normalized payoff parameters, G_n and F_n . Therefore, it is in turn the characteristics of a population, $F(\theta^i)$, and the material environment in which they interact, $\pi = (T, R, P, S)$ that determine whether or not a cooperative equilibrium is possible. The cooperating subpopulation is always a subset of the group of the *Assurance*-type players, since the *PD*-type players would never cooperate in a simultaneous game. The subset could include all the *Assurance*-type players, only a proportion of them, or none of them. The empty set of cooperators means that a cooperative equilibrium does not exist.

Divide the whole population Θ into two subsets: Θ^A of the players with *Assurance* preference and Θ^P of the players with *PD* preference. If $\Theta^c = \Theta^A$, an equilibrium exists in which all the players with *Assurance* preference cooperate. If $\Theta^c \subsetneq \Theta^A$, an equilibrium exists in which only a subset of the players with *Assurance* preference cooperate. If Θ^c is empty, there exists no cooperative equilibrium even when a relatively large proportion of the players is of *Assurance*-type. An example for each case is provided below. In these examples Γ is a game defined by its material payoff structure $\pi = (T, R, P, S : F_n, G_n)$, and the characteristics of a population are expressed in terms of the distribution of types: $\Theta[F(\theta^i) : F(\alpha^i, \beta^i)]$.

Example 16 $\Gamma(\pi, \Theta)$

$$\pi : F_n = G_n = 0.2$$

$$\Theta : F(\beta^i) = \beta \text{ and } \alpha^i = 1.1\beta^i$$

$F(\beta^i) = \beta$ implies that β^i is uniformly distributed over the range $[0,1)$. Then $\Pr(\beta^i \geq G_n) = 0.8$. Given π and Θ of this example, the probability that a player i is of *Assurance* type is 0.8. In spite of this quite high probability of a player having an *Assurance* preference, there exists no equilibrium in which some of the *Assurance* type players cooperate. To verify this, first see if there exists an equilibrium in which all of the *Assurance* preference type players cooperate. We can do this by comparing the expected utility of *Cooperation* and *Defection* for the players with *Assurance* preferences.

When a player with *Assurance* preference cooperates, there is a probability of 0.8 that the material payoff for him and his partner is (R, R) and a probability of 0.2 that he will end up with a material payoff (S, T) for him and his partner. Then the expected utility of cooperation for a player with *Assurance* preference, when all of the *Assurance*-type players cooperate, is

$$u^i(C) = 0.8(R) + 0.2(S - \alpha^i(T - S)). \quad (4.36)$$

Likewise, the expected payoff of *Defection* is

$$u^i(D) = 0.8(T - \beta^i(T - S)) + 0.2(P). \quad (4.37)$$

It is rational for player i to cooperate if and only if $u^i(C) \geq u^i(D)$, or

$$0.8(R) + 0.2(S - \alpha^i(T - S)) \geq 0.8(T - \beta^i(T - S)) + 0.2(P). \quad (4.38)$$

Substituting $1.1\beta^i$ for α^i we have

$$0.8[(R - T) + \beta^i(T - S)] \geq 0.2[(P - S) + 1.1\beta^i(T - S)]. \quad (4.39)$$

Dividing both sides by $T - S$ and multiplying both sides by 10, we get

$$8 \frac{-(T - R)}{(T - S)} + 8\beta^i \geq 2 \frac{P - S}{T - S} + 2.2\beta^i. \quad (4.40)$$

In inequality (4.40), $\frac{T-R}{T-S}$ is nothing but G_n , which is given as 0.2 in Γ of this example. Likewise, $\frac{P-S}{T-S} = F_n = 0.2$. Then the inequality can be simplified as

$$\begin{aligned} 5.8\beta^i &\geq 8(0.2) + 2(0.2) \\ \beta^i &\geq 2/5.8 = 0.34483. \end{aligned} \quad (4.41)$$

We can also check this in a simpler way by using the formula

$$\frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i} \leq \Pr(\theta^i \in \Theta^c | \theta^i \in \Theta), \quad (4.42)$$

which is a repetition of (4.32) in Condition 15.

$$\begin{aligned} \frac{0.2 + 1.1\beta^i}{\beta^i - 0.2 + 0.2 + 1.1\beta^i} &\leq 0.8 \\ \beta^i &\geq 2/5.8 = 0.34483 \end{aligned} \quad (4.43)$$

Substantively, (4.43) means that given $\Gamma = (\pi, \Theta)$ of this example, it is not enough for a player to have *Assurance* type preference even when all other *Assurance*-type players cooperate. β^i for the player should be larger than 0.34483 for him to rationally cooperate.

A kind of logical (but not temporal) cascade effect occurs leading to the elimination of any cooperative equilibrium. Now, if cooperation is ever possible, it should be done by the players for whom β^i is greater than 0.34483. Suppose that all the players with β^i greater than 0.34483 cooperate. In that case, the proportion of cooperators in the whole population is 0.65517. Or,

$$\Pr(i \in \Theta^c | i \in \Theta) = 0.65517. \quad (4.44)$$

The next step is to check whether it is rational for a player with β^i greater than 0.34483 to cooperate. Here again, we can check the possibility by verifying Condition 11.

$$\frac{0.2 + 1.1\beta^i}{\beta^i - 0.2 + 0.2 + 1.1\beta^i} \leq 0.65517$$

$$\beta^i \geq 0.72501. \quad (4.45)$$

Inequality (4.45) implies that only for the players with β^i greater than 0.72501, *Cooperation* is rational even when all the players with β^i greater than 0.34438 cooperate. In other words, $\Theta^c(\beta^i \geq 0.34438)$ is not a self-sustainable cooperative subset of Θ . Extending this line of analysis, it is not difficult to check that within $\Gamma = (\pi, \Theta)$ of Example 16, there is no sustainable Θ^c that satisfies the conditions of Proposition 14. Now we turn to another example.

Example 17 $\Gamma(\pi, \Theta)$

$$\begin{aligned}
\pi : \quad & F_n = G_n = 0.2 \\
\Theta : \quad & F(\beta^i) = 0.2 \quad \text{if } \beta^i < 0.2 \\
& = \frac{4}{9} + \frac{5}{9}\beta^i \quad \text{if } \beta^i \geq 0.2 \\
& \alpha^i = 1.1\beta^i
\end{aligned}$$

In this example, there is a 0.2 probability that a player i is strictly self-interested, and 0.8 probability that player i is of *Assurance* preference type. Among the *Assurance* preference type players, β^i is uniformly distributed over an interval $(0.5, 1)$. If all *Assurance*-type players cooperate, $\Pr(i \in \Theta^c | i \in \Theta) = 0.8$. All we need to do is to check whether *Cooperation* is rational for the player with the smallest β^i in Θ^c . Again, we can simply check the condition $\frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i} \leq \Pr(\theta^i \in \Theta^c | \theta^i \in \Theta)$ for the player with $\beta^i = 0.5$. If the inequality is satisfied, it can be said that it really is rational for the player to cooperate. And since it is rational to cooperate for the player with the least magnitude of inequality aversion, it is automatically rational for all others with the degree of inequity aversion bigger than him. So the subgroup of cooperation, Θ^c is self-sustainable and there exists a cooperative equilibrium in which all the *Assurance* preference type players do cooperate.

$$\begin{aligned}
\frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i} &\leq \Pr(\theta^i \in \Theta^c | \theta^i \in \Theta) \\
\frac{0.2 + 1.1(\beta^i)}{\beta^i - 0.2 + 0.2 + 1.1\beta^i} &\leq 0.8 \\
\beta^i &\geq 0.34483. \tag{4.46}
\end{aligned}$$

Since $\beta^i = 0.5$ for the player is bigger than 0.34483, the condition is met and there exists a cooperative equilibrium in which all of the *Assurance* preference type players cooperate. A final example is a case in which only a subset of *Assurance*-type players cooperate in a cooperative equilibrium.

Example 18 $\Gamma(\pi, \Theta)$

$$\pi : F_n = G_n = 0.2$$

$$\Theta : \Pr(\beta^i = 0) = 0.2$$

$$\Pr(\beta^i = 0.3) = 0.1$$

$$\Pr(\beta^i = 0.8) = 0.7$$

$$\alpha^i = 1.1\beta^i$$

In this example, 20% of individuals in the population Θ are strictly self-interested. Among the other 80% of players with *Assurance* preference, 10% have $\beta^i = 0.3$, and the other 70% have $\beta^i = 0.8$. If all of the players with *Assurance* preference cooperate, $\Pr(\theta^i \in \Theta^c | \theta^i \in \Theta) = 0.8$. Let us first see if it is rational for the players with $\beta^i = 0.3$ to cooperate when all of the *Assurance* preference type players cooperate. The procedure is the same as in Example 17, thus, β^i has to be greater than 0.344. Therefore, a proportion of *Assurance* preference type players (with $\beta^i = 0.3$) do not cooperate. The next step is to see if the remaining *Assurance* preference players still want to cooperate. Since players with $\beta^i = 0.3$ do not cooperate, $\Pr(\theta^i \in \Theta^c | \theta^i \in \Theta)$ is now 0.7.

$$\frac{F_n + \alpha^i}{\beta^i - G_n + F_n + \alpha^i} \leq \Pr(i \in \Theta^c | i \in \Theta)$$

$$\frac{0.2 + 1.1(\beta^i)}{\beta^i - 0.2 + 0.2 + 1.1\beta^i} \leq 0.7$$

$$\beta^i \geq 0.54054. \quad (4.47)$$

Since all the players in Θ^c have β^i equal to 0.7, the condition is met and there exists a cooperative equilibrium in which a subset of *Assurance* preference type players do cooperate.

In other words,

$$\Theta^c \neq \phi$$

$$\Theta^c \subsetneq \Theta^A.$$

The above analyses and examples suggest that the aggregate social outcomes are the results of the interaction between the characteristics of a population $\Theta(F(\theta^i))$ and the material conditions of the action situation $\pi(F_n, G_n)$. It is true, to an extent, that an individual's action depends on his own type (moral, ethical, or social inclination) expressed in terms of his preferences over the possible ways in which allocation of resources within a population can be made. However, his action also depends on the nature of the population within which he interacts with others and the material environment of the interaction. An individual with quite a strong propensity of equity and reciprocity may not cooperate when the incentives of defection are too large and there are too many individuals within the population who are not prepared to cooperate.

Equilibrium of Sequential Game

In addition to the structure of material payoffs and characteristics of population, institutions also matter in determining aggregate social outcomes. Here, we examine the impacts of the sequence of play by comparing equilibria of the sequential 2×2 social dilemma game with those of the simultaneous game. In general, when the game is played sequentially, the possibility of achieving mutual cooperation increases within a reasonable range of game parameters and type distributions. Strictly self-interested types cooperate when they are first movers under certain conditions of π and Θ . *Assurance*-type players always cooperate when they are the second movers and the first mover has cooperated.

Proposition 19 *Equilibrium of a sequential 2×2 social dilemma game:*

Let s^1 and s^2 denote strategies of the first and second mover, respectively.

$$\begin{aligned} s^1(\theta^1) &= C && \text{if } \frac{F_n + \alpha^1}{1 - G_n + \alpha^1} \geq \Pr(\theta^i \in \Theta^A | \theta^i \in \Theta) \\ &= D && \text{otherwise,} \end{aligned}$$

$$\begin{aligned} (s^2(\theta^2) | s^1 = C) &= C && \text{if } \beta^2 \geq G_n \\ &= D && \text{otherwise} \end{aligned}$$

$$(s^2(\theta^2) | s^1 = D) = D \quad \text{always}$$

is an equilibrium.

Proof. The second mover's strategy is intuitive. Since no player has a type of preference with which one cooperates even when the other player defects, the players always defect as a second mover when the first mover defects. When the first mover cooperates, the strategy of the second mover is strictly dictated by his preference ordering type. All

the *Assurance*-type players cooperate and all the *PD*-type players defect. Therefore, the key point of the proof is the strategy of the first mover, $s^1(\theta^1)$. This can be calculated by backwards induction, taking into account the strategy of the second mover. The first mover cooperates if, and only if, the expected utility of *Cooperation* is higher than that of *Defection*. When the first mover cooperates, there is $\mu^A (= \Pr(i \in \Theta^A | i \in \Theta))$ probability – the proportion of *Assurance*-type players in the population – of the outcome being (R, R) and $(1 - \mu^A)$ probability of the outcome being (S, T) . Therefore, the expected utility of *Cooperation* for the first player is

$$\begin{aligned}
 E[u^1(C|\theta^1)] &= \mu^A u^1(R, R) + (1 - \mu^A) u^1(S, T) \\
 &= \mu^A R + (1 - \mu^A)(S - \alpha^1(T - S)) \\
 &= \mu^A R + S - \mu^A S - \alpha^1(T - S) + \mu^A \alpha^1(T - S). \quad (4.48)
 \end{aligned}$$

Since all types of second movers *Defect* when the first mover defects, The expected utility of *Defection* is simply P .

$$E[u^1(D|\theta^1)] = P.$$

The first mover cooperates if, and only if, (4.48) is greater than or equal to P .

$$\begin{aligned}
 \mu^A R + S - \mu^A S - \alpha^1(T - S) + \mu^A \alpha^1(T - S) &\geq P \\
 \mu^A(R - S) + \mu^A \alpha^1(T - S) &\geq P - S + \alpha^1(T - S) \\
 \mu^A((T - S) - (T - R)) + \mu^A \alpha^1(T - S) &\geq P - S + \alpha^1(T - S) \quad (4.49)
 \end{aligned}$$

Dividing both sides of inequality (4.49) by $(T - S)$, we have

$$\begin{aligned}\mu^A(1 - G_n) + \mu^A\alpha^1 &\geq F_n + \alpha^1 \\ \mu^A(1 - G_n + \alpha^1) &\geq F_n + \alpha^1 \\ \mu^A &\geq \frac{F_n + \alpha^1}{1 - G_n + \alpha^1}.\end{aligned}\tag{4.50}$$

■

The right side of the inequality (4.50) increases in α^1 . Or, it is the smallest for the purely self-interested players for whom α^1 is 0. The implication is that given a fixed distribution of types within a population, it is harder for inequality aversion-type players to cooperate as the first mover than it is for the purely self-interested players. This is because purely self-interested players do not have additional utility loss from becoming a sucker, while inequality aversion players do have the additional loss. Put differently, inequity aversion-type players are more fearful of being exploited; thus, for them to cooperate as a first mover a higher proportion of the second movers with *Assurance* preference is needed.⁷

However, overall, the sequential play enhances the possibility of mutual cooperation. While some of the *Assurance* preference-type players may not cooperate in a simultaneous game, they always cooperate as a second mover when the first mover cooperates. In addition, while no player with *PD* preference cooperates in the simultaneous game, they

⁷This can be considered as the first intuitive anomaly of the inequity aversion model. One way to deal with this potential paradox is to relax the common knowledge condition of standard non-cooperative game theory such that an individual's belief about others' motivations is in part the projection of one's own motivation. In that case, it can be shown that inequity aversion-type individuals are more likely to cooperate as a first mover.

may cooperate as a first mover of the sequential game.

4.3.2 Altruism Model

Equilibria of Simultaneous Game

The problem of analyzing equilibria of the 2×2 social dilemma game based on the model of altruism is that it allows too many equilibria depending on Θ and π . Since $\pi(F_n, G_n)$ decides which kinds of preference types are possible and which are not, the equilibria also have to be calculated separately based on π . Here, we analyze only equilibria of a game $\Gamma(\Theta, \pi)$, in which

$$F_n > G_n \tag{4.51}$$

holds. (See Table 4.2 for necessary conditions in π for certain types of preference in the altruism model.) Equilibria of the game in which the reverse of (4.51) is true can be calculated in a comparable way.

As we did during the equilibria analyses based on the inequity aversion model, we also assume here that $F(\theta^i)$ is common knowledge. Given the material payoff condition π of the game Γ , a player can have one of the three preference types shown in Table 4.5 depending on the magnitude of his altruism parameter θ^i .

PD and *Angel* types, respectively, have the dominant strategies of *Cooperation* and *Defection*. Therefore, the analysis of equilibria has to focus on the behavior of *Assurance* preference-type players. Let us divide the population Θ into three subsets: Θ^P for the *PD* preference-type players, Θ^A for the *Assurance* preference-type players, and

θ^i	Preference Ordering	Preference Type
$0 \leq \theta^i \leq \frac{F_n}{1-F_n}$	$(T, S) > (R, R) > (P, P) > (S, T)$	<i>PD</i>
$\frac{F_n}{1-F_n} \leq \theta^i \leq \frac{G_n}{1-G_n}$	$(R, R) > (T, S) > (P, P) > (S, T)$	<i>Assurance</i>
$\frac{G_n}{1-G_n} \leq \theta^i \leq 1$	$(R, R) > (T, S) > (S, T) > (P, P)$	<i>Angel</i>

Table 4.5: Preference Types: Altruism Model with $\pi (F_n > G_n)$

Θ^G for the *Angel*-type players. Also, denote their proportions in the whole population as

$$\mu^P = \Pr(i \in \Theta^P | i \in \Theta)$$

$$\mu^A = \Pr(i \in \Theta^A | i \in \Theta)$$

$$\mu^G = \Pr(i \in \Theta^G | i \in \Theta).$$

Unlike the inequity aversion model, a strategy profile in which all the types of players *Defect* is not an equilibrium since *Angel*-type players have a dominant strategy of *Cooperation*. The rational strategy for the *Assurance*-type players can be found by comparing their expected utilities from *Cooperation* and *Defection*. Let's first see if there exists an equilibrium (and under which conditions) in which all *Assurance*-type players cooperate. If all *Assurance* preference-type players cooperate, an individual i with *Assurance* preference has a μ^P probability of ending up with a pair of material payoffs (S, T) for himself and his partner when he cooperates, and $1 - \mu^P$ probability of ending up with a pair of material payoffs (R, R) . Given his utility function (4.1), the expected utility can be calculated as follows:

$$Eu^i(C) = \mu^P(S + \theta^i T) + (1 - \mu^P)(R + \theta^i R). \quad (4.52)$$

Likewise, if he *defects*, there is a μ^P probability of ending up with a pair of material payoffs (P, P) for himself and his partner when the partner cooperates, and a $1 - \mu^P$ probability of ending up with a pair of material payoffs (T, S) . The expected utility is

$$Eu^i(D) = \mu^P(P + \theta^i P) + (1 - \mu^P)(T + \theta^i S). \quad (4.53)$$

Then, it is rational for this *Assurance*-type player to cooperate if and only if

$$\mu^P(S + \theta^i T) + (1 - \mu^P)(R + \theta^i R) \geq \mu^P(P + \theta^i P) + (1 - \mu^P)(T + \theta^i S). \quad (4.54)$$

Or,

$$\begin{aligned} \theta^i &\geq \frac{-\mu^P S - R + R\mu^P + \mu^P P + T - \mu^P T}{\mu^P T + R - R\mu^P - \mu^P P - S + \mu^P S} \\ \theta^i &\geq \frac{\mu^P(P - S) + T - R - \mu^P(T - R)}{\mu^P(T - R) + R - S - \mu^P(P - S)} \\ \theta^i &\geq \frac{\mu^P(P - S) + T - R - \mu^P(T - R)}{\mu^P(T - R) + T - T + R - S - \mu^P(P - S)} \\ \theta^i &\geq \frac{\mu^P(P - S) + T - R - \mu^P(T - R)}{\mu^P(T - R) + T - S - (T - R) - \mu^P(P - S)}. \end{aligned} \quad (4.55)$$

Dividing both the numerator and denominator of the right hand side of inequality (4.55)

by $(T - S)$, we have

$$\begin{aligned}\theta^i &\geq \frac{\mu^P F_n + G_n - \mu^P G_n}{\mu^P G_n + 1 - G_n - \mu^P F_n} \\ \theta^i &\geq \frac{\mu^P (F_n - G_n) + G_n}{\mu^P (G_n - F_n) + 1 - G_n}.\end{aligned}\quad (4.56)$$

In sum, in addition to the *Angel* preference-type players, a proportion of *Assurance* preference-type players with an altruism parameter θ^i greater than $\frac{\mu^P (F_n - G_n) + G_n}{\mu^P (G_n - F_n) + 1 - G_n}$ is willing to cooperate when all the *Assurance* type players cooperate. But the flip side of (4.56) is that a proportion of *Assurance*-type players with altruism parameter θ^i smaller than $\frac{\mu^P (F_n - G_n) + G_n}{\mu^P (G_n - F_n) + 1 - G_n}$ does not have the incentive for *Cooperation* even when all other *Assurance*-type players cooperate. Therefore, only when (4.56) holds for all the *Assurance*-type preference players, is it possible to have an equilibrium in which all the *Assurance*-type players cooperate.

What would happen if some of the *Assurance*-type players have θ^i smaller than $\frac{\mu^P (F_n - G_n) + G_n}{\mu^P (G_n - F_n) + 1 - G_n}$? The analysis resembles that regarding a self-sustaining subset of *Assurance*-type players in a cooperative equilibrium based on the inequity aversion model.

Proposition 20 *If there exists a subset Θ^{AC} of Θ^A such that for all $i \in \Theta^{AC}$ the following condition (4.57) holds, then there exists an equilibrium in which members of subset Θ^{AC} cooperate along with Angel-type players:*

$$\theta^i \geq \frac{(\mu^P + \mu^{AD})(F_n - G_n) + G_n}{(\mu^P + \mu^{AD})(G_n - F_n) + 1 - G_n} \quad (4.57)$$

where

$$\mu^{AD} = \mu^A - \mu^{AC}. \quad (4.58)$$

Proof. Notice that (4.57) is the same as (4.56) except that μ^P in (4.56) is replaced by $(\mu^P + \mu^{AD})$. Both μ^P in (4.56) and $(\mu^P + \mu^{AD})$ in (4.57) are the proportions of defectors in the respective equilibrium. While (4.56) is a condition for a self-sustaining equilibrium in which all the *Assurance* preference-type players cooperate, (4.57) is a generalized condition for any equilibrium in which a subset of *Assurance* preference players cooperate. Therefore, the proof of the proposition is similar to the procedures leading to (4.55) and (4.56). ■

Also notice that

$$\frac{(\mu^P + \mu^{AD})(F_n - G_n) + G_n}{(\mu^P + \mu^{AD})(G_n - F_n) + 1 - G_n} > \frac{G_n}{1 - G_n}, \quad (4.59)$$

because, original restriction (4.51), which is repeated as (4.60) below, prohibits *Chicken* type preferences.

$$G_n - F_n < 0. \quad (4.60)$$

The importance of (4.59) is as follows. $\frac{(\mu^P + \mu^{AD})(F_n - G_n) + G_n}{(\mu^P + \mu^{AD})(G_n - F_n) + 1 - G_n}$ is the reservation probability – called θ^* heretofore, – with which an *Assurance* preference-type player cooperates. Then, θ^* is the demarcation by which the *Assurance*-type players divide into a group of more altruistic and a group of less altruistic players. The former cooperates and the latter defects. On the other hand, $\frac{G_n}{1 - G_n}$ is the threshold for holding *Assurance* preferences in the first place. Therefore, it is natural that the threshold for cooperation, θ^* , is

θ^i	$0 \leq \theta^i < \frac{G_n}{1-G_n}$	$\frac{G_n}{1-G_n} \leq \theta^i < \theta^*$	$\theta^* \leq \theta^i < \frac{F_n}{1-F_n}$	$\frac{F_n}{1-F_n} \leq \theta^i \leq 1$
Type	<i>PD</i>	<i>Assurance</i>	<i>Assurance</i>	<i>Angel</i>
Strategy	<i>Defection</i>	<i>Defection</i>	<i>Cooperation</i>	<i>Cooperation</i>
Proportion	μ^P	μ^{AD}	μ^{AC}	μ^G
	$\theta^* = \frac{(\mu^P + \mu^{AD})(F_n - G_n) + G_n}{(\mu^P + \mu^{AD})(G_n - F_n) + 1 - G_n}$			

Table 4.6: Equilibrium of a 2×2 Social Dilemma Game : Altruism Model with $\pi : (F_n < G_n)$

greater than the threshold for *Assurance* preference. The whole population Θ divides into two subsets in the equilibrium. The cooperators' set includes Θ^G of the *Angel*-type players and a proportion of *Assurance*-type players Θ^{AC} . The defectors' subset of Θ includes Θ^P of strictly self-interested players and the remaining portion of the *Assurance* type players Θ^{AD} . Table 4.6 sums up the analysis.

We are now ready to sum up the properties of cooperative equilibrium, defined as an equilibrium in which at least a subset of the population Θ cooperates. Insofar as there exists at least one player with the altruism parameter θ^i greater than or equal to $\frac{G_n}{1-G_n}$, there always exists an equilibrium in which a subset of players Θ^C ($\Theta^G \subset \Theta^C \subset \Theta$) cooperates. In fact, since *Cooperation* is the dominant strategy of *Angel*-type players no matter what the distribution of other types, [*all Defection*] can never be an equilibrium when the *Angel*-type exists. The existence of *Angel*-type players is determined by $\Gamma = (\pi, \Theta)$.

The critical parameter threshold for cooperation θ^* increases in G_n

$$\frac{d}{dG_n} \theta^* = \frac{1 - (\mu^P + \mu^{AD})}{(-\mu G + \mu F - cG + cF - 1 + G)^2} > 0, \quad (4.61)$$

suggesting that, consistent with our intuition, it becomes harder for an *Assurance*-type player to join the cooperators' group as the size of *Greed* becomes larger. Also consistent

with our intuition, the larger the proportion of *PD* preference-type players, the higher is the threshold θ^* and it becomes harder for an *Assurance*-type player to join the cooperators' group.

$$\frac{d}{d\mu^P}\theta^* = \frac{F_n - G_n}{(\mu^P G_n - \mu^P F_n + \mu^{AD} G_n - \mu^{AD} F_n + 1 - G_n)^2} > 0. \quad (4.62)$$

Equilibria of Sequential Game

To make the equilibrium analysis comparable to that of the simultaneous game, the structure of material payoffs π is restricted such that $F_n > G_n$. The substantive implication is that *Chicken* types do not exist in the population Θ .

The subgame perfect equilibrium can be calculated by first looking at the rational behavior of the second movers. When the first mover cooperates, both the *Angel* and *Assurance* types reciprocate by cooperating in return, but *PD* type players do not. When the first mover defects, only the *Angel* type cooperates, while *PD* and *Assurance* types defect in return. At the first mover's position, *Angel*-type players always cooperate since that is their dominant strategy regardless of the strategy of the second mover. For the *PD* and *Assurance*-type players as the first movers, their strategies depend on their expectations of the strategy by the second movers which, in turn, can be calculated from the original distribution of types within the population. Therefore, the focus of analysis is the choice of the first mover with *Assurance* or *PD* type. Let us start with *PD* types. The expected utility of *Cooperation* for *PD*-type first movers, given the contingent strategy of each type of second movers and the distribution of types within the population, is:

$$u^i(C) = \mu^P(S + \theta^i T) + (1 - \mu^P)(R + \theta^i R). \quad (4.63)$$

This is because when he cooperates at the first mover's position, there is μ^P probability (the proportion of *PD*-type players in the population) of becoming a *Sucker*, while there is $(1 - \mu^P)$ probability (the sum of the proportion of the *Assurance* and *Angel*-type players within the population) of being reciprocated.

The expected utility of *Defection* for him is

$$u^i(D) = (1 - \mu^G)(P + \theta^i P) + \mu^G(T + \theta^i S), \quad (4.64)$$

because when he defects, only the *Angel*-type players continue to cooperate while both the *PD* and *Assurance*-type players defect in return. A *PD*-type player cooperates as a first mover if, and only if, (4.63) is greater than or equal to (4.64), or

$$\mu^P(S + \theta^i T) + (1 - \mu^P)(R + \theta^i R) \geq (1 - \mu^G)(P + \theta^i P) + \mu^G(T + \theta^i S). \quad (4.65)$$

After simplifying the inequality by rearranging the arguments and dividing both sides by the normalization factor $(T - S)$, we have

$$\theta^i \geq \frac{\mu^P(1 - G_n) - C_n + \mu^G(1 - F_n)}{\mu^P G_n + C_n + \mu^G F_n}. \quad (4.66)$$

The right side of (4.66) – denoted as θ^* heretofore – is the threshold for players other than the *Angel* type to cooperate as a first mover of a sequential 2×2 social dilemma game. Then, the equilibrium of the sequential 2×2 social dilemma game based on the model of altruism can be formalized.

Proposition 21 *Equilibrium of a sequential 2×2 social dilemma game: Let s^1 and s^2 denote strategy of the first and second mover, respectively. Also define $\theta^* = \frac{\mu^P(1-G_n) - C_n + \mu^G(1-F_n)}{\mu^P G_n + C_n + \mu^G F_n}$.*

Then,

$$\begin{aligned}
 s^1(\theta^1) &= C & \text{if } \theta^1 \geq \min\left(\frac{F_n}{1-F_n}, \theta^*\right) \\
 &= D & \text{otherwise} \\
 (s^2(\theta^2)|s^1 = C) &= C & \text{if } \theta^2 \geq \frac{G_n}{1-G_n} \\
 &= D & \text{otherwise} \\
 (s^2(\theta^2)|s^1 = D) &= C & \text{if } \theta^2 \geq \frac{F_n}{1-F_n} \\
 &= D & \text{otherwise}
 \end{aligned}$$

is the unique equilibrium of a sequential 2×2 social dilemma game. That is: (1) among the first movers, all the Angel-type players and a subset of the PD type and Assurance type - which could be either the empty set or the whole set, or any subset in between - cooperate; and (2) among the second movers, the Angel-type players always Cooperate, the Assurance-type players always reciprocate, and the PD-type players always Defect.

The threshold parameter value θ^* is determined by the material payoff structure $\pi = (T, R, P, S)$ and the distribution of types within the population $F(\theta^i)$. The larger the material gain from mutual *Cooperation*, the smaller the threshold and the more likely a first mover with either *PD* or *Assurance*-type preference will cooperate.

$$\frac{d}{dC_n} \theta^* = -\frac{\mu^P + \mu^G}{(\mu^P G_n + C_n + \mu^G F_n)^2} < 0. \quad (4.67)$$

The larger the proportion of the *PD*-type players within the population, the higher the

threshold and the less likely a first mover with either *PD* or *Assurance*-type preference will cooperate.

$$\frac{d}{d\mu^P}\theta^* = \frac{C_n + \mu^G(F_n - G_n)}{(\mu^P G_n + C_n + \mu^G F_n)^2} > 0. \quad (4.68)$$

On the other hand, the proportion of the *Angel*-type players does not always have a positive impact on the likelihood of cooperation by the first movers.

$$\frac{d}{d\mu^G}\theta^* = \frac{C_n - \mu^P(F_n - G_n)}{(\mu^P G_n + C_n + \mu^G F_n)^2}. \quad (4.69)$$

The sign of the right side of (4.69) is indeterminate, suggesting that the impact of the proportion of the *Angel*-type players depends on the exact value of other parameters. The impact is indeterminate, because the presence of the *Angel* type players gives additional incentive to defect for the *PD*-type first players.

4.4 Behavior in Four Information Sets

While the traditional self-interest model of a 2×2 social dilemma presumes that all individuals will defect in both the simultaneous and sequential games, the models of altruism and inequity aversion predict that cooperation is possible under certain conditions. Since the conditions involve the distribution of types, which are not directly observable, predictions cannot be made in the form of point predictions – the exact proportion of cooperators and defectors.

The strategy of the empirical tests adopted in this study is to derive, from the

equilibria analyses of the previous section, implications of the models about the *relative* frequencies of *Cooperation* in the four distinct information sets of the simultaneous and sequential 2×2 social dilemma games. In the following, it is assumed that players follow a cooperative equilibrium whenever one exists. The four qualitatively different information sets are:

1. two information sets in the simultaneous game;
2. first mover's information set in the sequential game;
3. second mover's information set in the sequential game following first mover's *Cooperation*;
and
4. second mover's information set in the sequential game following first mover's *Defection*.

Let us denote the frequencies of *Cooperation* at the four information sets $\Pr(C|I_{Sm})$, $\Pr(C|I_{Sq1})$, $\Pr(C|I_{Sq2,C})$, and $\Pr(C|I_{Sq2,D})$, respectively.

4.4.1 Inequity Aversion Model

The inequity aversion model predicts, first of all, that no individual playing a 2×2 social dilemma as a second mover would cooperate knowing that the first mover has defected. This result is formalized in Hypothesis $H_{IA} - 1$.⁸

$$H_{IA} - 1 : \Pr(C|I_{Sq2,D}) = 0 \quad (4.70)$$

⁸ $H_{IA} - 1$ means a first(1) hypothesis(H) derived from the inequity aversion model (IA). In the following, the same notational convention will be used.

$H_{IA} - 1$ is an important result of the inequity aversion utility function. For example, if the nature of non-selfishness is to be captured by the relative weights a player attaches to one's own and the other player's material well-being (as the model of altruism does), we have to see some of the second movers in the sequential dilemma games cooperating even when the first mover defects. Or, if the extent to which a player receives utility from conducting morally correct action is the correct dimension on which to describe interindividual heterogeneity, again we would see some of the second movers cooperating even when the first move has defected. In that sense, $H_{IA} - 1$ is a strong prediction.

The frequency of *Cooperation* among the players in the other three information sets is at least as high as that among the second movers of the sequential game when the first mover has defected. Specifically, let us first compare, assuming that the population characteristics are the same among different conditions, $\Pr(C|I_{Sm})$ and $\Pr(C|I_{Sq2,C})$, the behavior of the players in the simultaneous game and that of the second movers in the sequential dilemma game when the first mover has cooperated. The equilibrium analyses (Proposition 10) in the previous section suggest that all of the *Assurance* preference-type second movers cooperate following first movers' *Cooperation*. In the simultaneous game as Proposition 14 suggests, only a subset (which could be an empty set or the whole set itself) of the *Assurance*-type players cooperate. We can develop another hypothesis that incorporates this result and $H_{IA} - 1$.

$$H_{IA} - 2 : \Pr(C|I_{Sq2,C}) \geq \Pr(C|I_{Sm}) \geq \Pr(C|I_{Sq2,D}) \quad (4.71)$$

Notice that $H_{IA} - 2$ applies to all possible $\Gamma = (\pi, \Theta)$ – meaning that it is true, regardless

of the structure of the material payoffs and the distribution of types within a population. However, $\Pr(C|I_{Sq1})$, the frequency of cooperation among the first movers of the sequential game and its relative standing compared to $\Pr(C|I_{Sq2,C})$ and $\Pr(C|I_{Sm})$ cannot be generally determined. It varies, depending upon the distribution of types and the material payoff structure. We can say only that the frequency of cooperation among the first movers of a sequential game is at least as high as that among the second movers of the same sequential game when the first mover has defected.

$$H_{IA} - 3 : \Pr(C|I_{Sq1}) \geq \Pr(C|I_{Sq2,D}) \quad (4.72)$$

4.4.2 Altruism Model

For the purpose of comparison, let us see if the hypotheses derived based on the inequity aversion model also hold in the altruism model. First, is it possible to observe cooperation by the second movers of the sequential game when the first mover has defected? In other words, can $\Pr(C|I_{Sq2,D})$ be greater than 0? The proportion of cooperators among the second movers of a sequential game when the first mover has defected is nothing but the proportion of *Angel* preference types within the population, or μ^G . The exact value of μ^G depends on π and Θ . Therefore, all we can say regarding $\Pr(C|I_{Sq2,D})$ in the altruism model is

$$H_{AT} - 1 : \Pr(C|I_{Sq2,D}) \geq 0. \quad (4.73)$$

$H_{AT} - 1$, in fact, is not a hypothesis at all, since it does not restrict the possible

world in any sense. We turn to the second hypothesis based on the inequity aversion model to see if anything comparable exists in the altruism model. The essence of $H_{IA} - 2$ was that one should observe at least as many cooperators among the second movers of the sequential game given that the first mover has cooperated as those in the simultaneous game. The latter in turn is at least as many as those among the second movers of the sequential game when the first mover has defected. This also holds in the altruism model. The reasons are as follows. First, while all the *Assurance* and *Angel*-type players cooperate when they are the second movers of the sequential game given that the first mover has cooperated, only a subset of *Assurance*-type players cooperates along with *Angel* types when they play the simultaneous game. Second, while only the *Angel*-type second movers cooperate in the sequential game when the first mover has defected, a subset of *Assurance*-type players also cooperates in the simultaneous game. Therefore,

$$H_{AT} - 2 : \Pr(C|I_{Sq2,C}) \geq \Pr(C|I_{Sm}) \geq \Pr(C|I_{Sq2,D}) \implies H_{IA} - 2. \quad (4.74)$$

Turning to the third hypothesis derived from the inequity aversion model, we again notice that the proportion of cooperators among the first movers of the sequential game is underrestricted with only one qualification: that it is at least as big as that among the second movers of the sequential game when the first mover has defected. This is because while only the *Angel*-type players cooperate as the second movers of the sequential game given that the first mover has defected, a proportion of players in the other subset, composed of the *PD* and *Assurance* types, also cooperates in the simultaneous game depending on

π and $F(\theta^i)$. This result, in turn, is the same as that based on the inequity aversion model:

$$H_{AT} - 3 : \Pr(C|I_{Sq1}) \geq \Pr(C|I_{Sq2,D}) \implies H_{IA} - 3. \quad (4.75)$$

We notice that the only difference between the two models regarding behavior in the four different information sets is their respective prediction regarding behavior of the second movers in the sequential game when the first mover has defected. However, the test is rather difficult because the relevant hypothesis based on altruism does not contain restriction. We can say only that if $\Pr(C|I_{Sq2,D})$ truly is 0, the inequity aversion model is confirmed in that regard, while the altruism model is simply not falsified by virtue of its nonfalsifiability.

4.5 Empirical Tests and Results

4.5.1 Data

Before examining the experimental results, this subsection briefly describes the experimental procedure used to generate AOW and SURVEY data.

AOW: Participants were recruited from introductory and intermediate-level economics classes at Indiana University. The total number of participants was 166 (86 males and 80 females). Monetary incentives were emphasized in the recruitment. Subjects were informed publicly that they would participate in a decision problem that would be played only once. Double blind procedures were used. That is, decisions were anonymous to other participants and to the experimenter. The anonymity of decisions was assured by the use

		<i>Individual 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Individual 1</i>	<i>Cooperation</i>	\$10, \$10	\$0, \$15
	<i>Defection</i>	\$15, \$0	\$5, \$5

Figure 4.6: Monetary Incentive Structure of AOW Experiment

of two experimenters and two room monitors. The role of one experimenter was to interact with the participants, while the other experimenter processed participants' decisions.

The 2×2 social dilemma game used in the experiment was constructed in the following way. Each participant was promised \$5 by the experimenter and was then asked to decide whether or not to give that \$5 to their partner (a person randomly chosen from the participants in the other room). When the participant gave \$5, the partner received \$10. When the participant did not give \$5, he/she could keep the \$5 plus another \$10 if his/her partner chose to give his/her \$5. Therefore, both participants received \$5 when both decided not to give \$5, and both received \$10 when both decided to give \$5. When one of the two participants decided not to give \$5 and the other participant decided to give \$5, the former received \$15 and the latter received nothing. In addition to game payoffs, all participants received a \$5 "show-up" fee. Figure 4.6 shows the incentive structure of the 2SD game while referring "Giving \$5" to be *Cooperation* and "Not Giving \$5" to be *Defection*.

Participants were recruited to two different classrooms. Upon arrival at the classroom, participants were seated and received an opaque envelope that contained two ID cards and a decision sheet. Envelopes were distributed randomly so that the experimenter had no way of matching participants with their ID cards. Ten to thirty subjects participated in any one experimental session.

After all participants arrived at the laboratory, the experimenter gave the following oral instructions. (1) This is an experiment in decision making. (2) The experiment takes place in two rooms; earnings depend on the individual's decision in the experimental game and on that of a randomly chosen person in the other room. (3) Money earned will be paid in cash at the end of the experiment and all decisions and earnings are anonymous. (4) Participants could ask questions at any time.

Following introductory announcements, participants read the written instructions explaining the procedures and the nature of the decision game. Participants then answered four questions about the game to confirm their understanding. The experimenter then reviewed the decision-making game and the answers to the four questions.

Participants made their decision by completing the decision form. The decision form and one of the participants' ID cards were then put in an envelope by the participant and picked up by the experimenter. Participants kept one ID card to identify themselves for payoffs. Participants then answered a postexperimental questionnaire. In the conditions requiring sequential decision making, the experimenter collected the decisions from the room with the first movers, then went to the room with the second movers. The second movers received a form that included the decision that the partner had already made.

When all participants finished answering the postexperimental questionnaire, they received their earnings for the experiment by showing their ID card to the monitor and receiving a closed envelope identified with their ID number. Each experimental session took approximately 45 minutes.

SURVEY: The Survey data is from a series of classroom surveys conducted during

	Other	
You	A	B
A	You: \$10.00 Other: \$10.00	You: \$25.00 Other: \$ 5.00
B	You: \$5.00 Other: \$25.00	You: \$20.00 Other: \$20.00

Figure 4.7: Decision Problem in SURVEY

the spring semester of 1999 in the three undergraduate courses (one introductory microeconomics course, one honors economics course, and one political science course) at Indiana University. Participation was voluntary; students in the courses were not required to answer the survey questionnaire as a requirement for the course. On a double-sided paper, an imaginary decision situation was presented. Students were asked to assume that they had to make a decision whose monetary outcome was affected by a similar decision made by another student in the class. The decision situation – a 2×2 social dilemma – was shown in a matrix that was exactly replicated in Figure 4.7.⁹

Students were first asked to check off one of the two boxes (I would choose A, I would choose B) below the matrix that he or she would choose if faced such an opportunity. And then the questionnaire also asked several more questions related to the satisfactory level of each of the four possible outcomes, their belief about others' choice, and their level of general trust.

4.5.2 Types of Preference Orderings

In both the experiments, subjects were asked to response to four questions related to their preferences regarding the four possible outcomes of a 2×2 social dilemma game.

⁹It is a 2×2 social dilemma with *Normalized Greed* = 0.25, *Normalized Fear* = 0.25, and *Normalized Cooperators' Gain* = 0.5. Choice A corresponds to *Defection* and B to *Cooperation*.

The questions in SURVEY are:

1. How satisfactory would it be to you if both you and the other player chose D?

1 ___ 2 ___ 3 ___ 4 ___ 5 ___ 6 ___ 7

Very Unsatisfactory

Very Satisfactory

2. How satisfactory would it be to you if both you and the other player chose C?

1 ___ 2 ___ 3 ___ 4 ___ 5 ___ 6 ___ 7

Very Unsatisfactory

Very Satisfactory

3. How satisfactory would it be to you if you chose D and the other player chose C?

1 ___ 2 ___ 3 ___ 4 ___ 5 ___ 6 ___ 7

Very Unsatisfactory

Very Satisfactory

4. How satisfactory would it be to you if you chose C and the other player chose D?

1 ___ 2 ___ 3 ___ 4 ___ 5 ___ 6 ___ 7

Very Unsatisfactory

Very Satisfactory

Questions in AOW have different wording because of the different framing of the action situation in the experiment. For example, a question in AOW that asks “How satisfactory would it be if both you and your partner gave \$5.00?” corresponds to the second question in SURVEY above. At this stage of analysis, the focus is strictly on the ordinal ranking of the four outcomes. Analyses of possible preference-ordering types in Section 2 intentionally ignored weak orderings – the possibility of indifference over two or more outcomes. Since the questions allow ties, our analyses needed to be modified accordingly. Tables 4.7 and 4.8 present preference-ordering possibilities including weak orderings for the

Type	Ordering
PD	$(T, S) > (R, R) > (P, P) > (S, T)$
Indifference (PD/Assurance)	$(T, S) = (R, R) > (P, P) > (S, T)$
Assurance	$(R, R) > (T, S) > (P, P) > (S, T)$
Assurance2	$(R, R) > (P, P) >^i (T, S) > (S, T)$

Table 4.7: Preference-Ordering Possibilities: Inequity Aversion Model

Type	Ordering
PD	$(T, S) > (R, R) > (P, P) > (S, T)$
Indifference 1 (PD/Chicken)	$(T, S) > (R, R) > (P, P) = (S, T)$
Indifference 2 (PD/Assurance)	$(T, S) = (R, R) > (P, P) > (S, T)$
Assurance	$(R, R) > (T, S) > (P, P) > (S, T)$
Assurance 2	$(R, R) > (P, P) > (T, S) > (S, T)$
Chicken	$(T, S) > (R, R) > (S, T) > (P, P)$
Angel	$(R, R) > (T, S) > (S, T) > (P, P)$
Indifference 3	$(R, R) = (T, S) > (P, P) = (S, T)$
Indifference 4 (Angel/Chicken)	$(R, R) = (T, S) > (S, T) > (P, P)$
Indifference 5 (Angel/Assurance)	$(R, R) > (T, S) > (S, T) = (P, P)$

Table 4.8: Preference-Ordering Possibilities: Altruism Model

inequity aversion model and the altruism model, respectively. Tables 4.9 and 4.10 present distribution of preference types in the two sets of data for each of the two models.

The results reported in Tables 4.9 and 4.10 show, first of all, that the common assumption of self-interest, which allows only the *PD* preference-ordering type, is not satisfactory at all. The *PD* preference type itself is the most frequently observed preference type in both the data sets. In SURVEY, almost 40% of the subjects revealed this type of preference. In AOW, the percentage is a bit lower, about 20%. However, in both the data sets, the self-interest model accounts for less than 50% and there are other preference types with significant proportions.

The inequity aversion model increases two more preference types (three more when indifference is allowed). The marginal increase in explained data by the three additional

Type	SURVEY	AOW99
PD	141 (39.7 %)	33 (19.8 %)
Indifference (PD/Assurance)	67 (18.9 %)	26 (15.6 %)
Assurance	16 (4.5 %)	17 (10.2 %)
Assurance 2	0 (0.0 %)	14 (8.4 %)
Not explained	131 (36.9 %)	77 (46.1 %)
Anomaly 1: $(S, T) > (T, S)$	40 (11.3 %)	10 (6.0 %)
Anomaly 2: $(R, R) = (P, P)$	30 (8.5 %)	24 (14.4%)
Total		167

Table 4.9: Distribution of Types: Inequity Aversion Model

Type	SURVEY	AOW99
PD	141 (39.7 %)	33 (19.8 %)
Indifference 1 (PD/Chicken)	4 (1.1 %)	3 (1.8 %)
Indifference 2 (PD/Assurance)	67 (18.9 %)	26 (15.6 %)
Assurance	16 (4.5 %)	17 (10.2 %)
Assurance2	0 (0.0 %)	14 (8.4 %)
Chicken	0 (0.0 %)	2 (1.2 %)
Angel	1 (0.3 %)	3 (1.8 %)
Indifference 3	10 (2.8 %)	3 (1.8 %)
Indifference 4 (Angel/Chicken)	0 (0.0 %)	3 (1.8 %)
Indifference 5 (Angel/Assurance)	2 (0.6 %)	3 (1.8 %)
Not explained	116 (32.8 %)	60 (39.3 %)
Anomaly 1: $(S, T) > (T, S)$	40 (11.3 %)	10 (6.0 %)
Anomaly 2: $(R, R) = (P, P)$	30 (8.5 %)	24 (14.4%)
Total		167

Table 4.10: Distribution of Types: Altruism Model

		SURVEY	AOW
Self-Interest	(1 type)	39.7 %	19.8 %
Inequity Aversion	(4 types)	23.4 %	34.0 %
Altruism	(10 types)	7.0 %	7.0 %

Table 4.11: Marginal Explanatory Power of Three Models

types is quite significant. In SURVEY, the inequity aversion model accounts for about 73%, while in AOW, it explains about 54%. Specifically, a significant proportion of subjects revealed in both SURVEY and AOW that they are indifferent between the two outcomes (T, S) and (R, R) ; this type of preference is the second largest in both data sets. Also present is the strict *Assurance* preference type: 4.51% in SURVEY and 19.5% in AOW.

The model of altruism expands the space of possible preferences significantly. However, the marginal explanatory power gained by the expansion is not quite significant. Six additional preference types, added by the model of altruism to the model of inequity aversion, explain only about 7% more in each of SURVEY and AOW. That is, addition of one more preference type on average only extends the explanatory power of the model by only 1% of the data. Table 4.11 shows the marginal explanatory power gained by the models of inequity aversion and that of altruism.

4.5.3 Behavior in Four Information Sets

The empirical test of the relative performance of the two models in terms of explaining behavior is not quite straightforward. First of all, a strict test of the models would require that the type – the exact value of type parameters – of each of the players be known. However, a player's type can be at best conjectured (or restricted within a certain range) based on his answer to the survey questionnaire. Survey data regarding preference ordering

Information Set	% <i>Cooperation</i>	# of Cooperators	# Observation
I_{Sm}	36%	37	104
I_{Sq1}	56%	18	32
$I_{Sq2,C}$	61%	11	18
$I_{Sq2,D}$	0%	0	13

I_{Sm} : Information set in the simultaneous game
 I_{Sq1} : First movers information set in the sequential game
 $I_{Sq2,C}$: Second mover's information set following first mover's *Cooperation*
 $I_{Sq2,D}$: Second mover's information set following first mover's *Defection*

Table 4.12: Frequency of *Cooperation* in Four Information Sets

are quite noisy. There are many ordering types that cannot be accounted by either of the models. In addition, since the models differ in terms of the method of distinguishing types, a common dimension of type-distribution does not exist.

However, a test of the models is not totally impossible. We adopt two criteria comparable to those used in evaluating the models' relative performance with regard to the allowed preference-ordering types. A model is better when it restricts the possible world in a meaningful way. A model with underrestriction may not be falsified by empirical evidence, but the underrestrictiveness of a model is its weakness rather than strength. Second, a model has to pass the empirical test. Restrictiveness of a theory is a necessary but not sufficient condition. A theory, however elegant, cannot be considered useful unless it survives empirical tests. This subsection conducts a test of the models based on their predictions regarding individuals' behavior in the four qualitatively different information sets of the simultaneous and sequential 2×2 social dilemma games. The qualitative differences among the four information sets are discussed and the hypotheses regarding the relative frequencies of *Cooperation* are derived in Section 4.

Table 4.12 presents relative frequency of cooperation in the four qualitatively dif-

ferent information sets of 2×2 social dilemma games. The results confirm all the hypotheses, (4.70) to (4.75), derived in Section 4. Of particular interest is the absence of *Cooperation* among the second movers of the sequential game when the first mover has defected. This result strongly favors the inequity aversion model.

Also interesting is a measure of the proportion of *Assurance/Angel* types. While behavior in the simultaneous game and the first mover's information set in the sequential game do not directly indicate a player's type, that in the second mover's information set in sequential game does. A choice of *Cooperation* by a second mover of the sequential game following the first mover's *Cooperation* indicates that his is of an *Assurance* preference type (inequity aversion model) or either an *Assurance* or *Angel* type (altruism model). Choice of *Cooperation* by a second mover in the sequential game following the first mover's *Defection* indicates that the individual is of an *Angel* type (altruism model) while the behavior is an anomaly in the model of inequity aversion. We can deduce from the third row of Table 4.12 that within the population of this specific experiment, assuming that the assignment of subjects to each information set was random, about 61% have non-selfish preferences.

Though the results strongly support the inequity aversion model over the model of altruism as a proper alternative to the model of pure self-interest, a single experiment is not enough to draw too strong a conclusion. Table 4.13 compares the current result with those of other experiments conducted independently. The compared experiments are the more interesting because of their use of subjects from different countries. The results of Cho and Choi (1999) also strongly support the model of inequity aversion. All of the hypotheses

Information Sets	Hayashi et al., 1999 (Japan)	Cho and Choi, 1999 (Korea)
I_{Sm}	56% (15/27)	46% (28/59)
I_{Sq1}	83% (19/23)	52% (11/21)
$I_{Sq2,C}$	75% (15/20)	73% (8/11)
$I_{Sq2,D}$	12% (3/25)	0% (0/10)

Table 4.13: Frequency of *Cooperation* in Four Information Sets: A Comparison of Studies

(4.70) to (4.75) regarding the relative frequencies of *Cooperation* are confirmed. In particular, no *Cooperation* was observed among ten Korean subjects who played a sequential 2×2 social dilemma game as second movers following the first movers' *Defection*. However, Hayashi et al.'s (1999) results regarding Japanese subjects' behavior reveal a different behavioral tendency. Three out of 25 subjects chose *Cooperation*, knowing that the first movers' had already chosen *Defection*. The possibility of an unconditionally cooperating type cannot be totally ignored, though their presence is rarely observed. On the other hand, the high proportion of reciprocators (75% in Hayashi et al. and 73% in Cho and Choi) who chose *Cooperation* knowing that the first movers had already chosen *Cooperation*, confirms the existence of a significant proportion of nonselfish preference-type players across different populations.

4.6 Conclusion

This chapter conducted a series of theoretical and empirical investigations of two alternative motivations to selfishness: altruism and inequity aversion. When applied to 2×2 social dilemma games, each of the two models generates a series of possible preference-ordering types over the four outcomes of the game, while the traditional assumption of self-interest allows only one preference type. The model of altruism classifies individuals

into four preference-ordering types, but the possibilities of certain types are limited by the structure of material payoffs of a game. On the other hand, the model of inequity aversion generates only two preference-ordering types and their existence is not limited by the structure of the material payoffs. The altruism model predicts that there are individuals who cooperate no matter what another individual does. The model of inequity aversion, on the other hand, precludes unconditional cooperation and divides individuals into two broad subsets of conditional cooperators and unconditional defectors.

When the social dilemma is framed as an incomplete information game, both the models specify conditions for cooperative equilibrium in simultaneous and sequential 2×2 social dilemma games. The equilibrium analysis allows us to derive hypotheses regarding the relative frequency of cooperation in the four qualitatively different information sets of the games. The hypotheses based on the altruism model are less restrictive than those that are based on the model of inequity aversion.

Empirical tests are conducted drawing on two sets of experimental data. In terms of preference ordering, the model of inequity aversion accounts for a substantive proportion of the preference types not explained by the pure selfishness model. In contrast, the altruism model does not provide meaningful additional explanation for the types that are not accounted by the inequity aversion model. In terms of the behavior in the four qualitatively different information sets, the data strongly supports the hypotheses based on the model of inequity aversion.

Chapter 5

Finite Repetition of a 2×2 Social Dilemma

5.1 Introduction

Chapter 4 developed a model of a 2×2 social dilemma game, of which the material payoff structure is the *Prisoner's Dilemma*, but with players of different types defined by the ways they transform material payoffs into von Neumann-Morgenstern utilities. There, a stage game was defined without abandoning any requirements necessary to define a game in standard game theory. Equilibria and supporting conditions were analyzed.

This chapter studies the finitely repeated 2×2 social dilemma game. In part, the subject is related to the study of the finitely repeated *Prisoner's Dilemma* game. However, if one understands the payoffs of a game in terms of von Neumann-Morgenstern utilities and applies the standard solution concepts, there is no way to show that *Cooperation* occurs in

the equilibrium of a single stage or a finitely repeated version of the *Prisoner's Dilemma* game.¹ Therefore, the most famous and most game-theoretic model of *Cooperation* in the finitely repeated *Prisoner's Dilemma* by Kreps et al. (1982) depends on either relaxation of the rationality and common knowledge assumptions in the standard game theory or a possibility that the material payoffs do not map into von Neumann-Morgenstern utilities in a one-to-one fashion. Kreps et al. and Fudenberg and Maskin (1986) show that if a rational player assesses a small probability that his partner is irrationally playing *Tit-for-Tat* strategy, he or she may cooperate except for the final few stages. However, the evidence in the field and laboratory indicates that the types of individuals whose behavior is best described as *reciprocal* deserve a better place in theoretical analyses than that of a hypothetical and irrational factor in the rational egoists' decision-making problem. Ostrom (1998: 4) notes:

To assume that if some players *irrationally* choose reciprocity, then others can *rationally* choose reciprocity is a convoluted explanation—to say the least—of the growing evidence that reciprocity is a core norm used by many individuals in social dilemma situations.

By formalizing non-selfish motivations in utility functions, this study models all individuals as rational decisionmakers in the sense that they have preferences and try to maximize their expected utility with action. In that regard, this study resembles the second model of Kreps et al. in which players have a common and objective prior regarding the probability that a player is of the *Assurance* type. Kreps et al. defer actual analyses of the sequential equilibria to the readers. In this chapter, the sequential equilibria of the finitely repeated 2×2 social dilemma are studied. The equilibrium analyses will help

¹There exist cooperative Nash equilibria of the finitely repeated *Prisoner's Dilemma* game, but those equilibria involve incredible threats or promises.

		<i>Individual 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Individual 1</i>	<i>Cooperation</i>	1, 1	b, a
	<i>Defection</i>	a, b	0, 0
		$a > 1, b < 0$	
		$a + b < 2$	

Figure 5.1: 2×2 Social Dilemma Action Situation

us to understand how the material payoff structure and the distribution of types within a population affect the possibility of cooperation and the dynamics that unfold when the 2×2 social dilemma game is repeated a finite number of times. The implications of equilibrium analyses will be tested empirically using an experimental data set. In the regression analyses of the data, special attention will be paid to the problem of interdependency between the strategy choices of any two paired players and that of the heterogeneity across individuals.

5.2 Finitely Repeated 2×2 Social Dilemma Game with Uncertainty

Cooperative equilibria of the finitely repeated 2×2 social dilemma games, when they exist, are diverse and complicated. To simplify formal analyses, this chapter introduces a different representation of the 2×2 social dilemma action situation than was used in the previous chapter. However, in modeling different motivations of individuals, the basic empirical conclusions of Chapter 4 will be maintained. That is, individuals are modeled to have either an *Assurance* or a *PD* preference over the four outcomes of a 2×2 social dilemma. Figure 5.1 shows the matrix representation of a 2×2 social dilemma that will be used throughout this chapter.

		<i>Individual 2</i>	
		<i>Cooperation</i>	<i>Defection</i>
<i>Individual 1</i>	<i>Cooperation</i>	1, 1	b, $a - \theta^2$
	<i>Defection</i>	$a - \theta^1$, b	0, 0

$a - \theta^1 > b \ (\Rightarrow \ \theta^1 < a - b)$

Figure 5.2: Von Neumann-Morgenstern Utilities in the 2×2 Social Dilemma Game

Figure 5.1 draws on Kreps et al.'s (1982) model of the stage *Prisoner's Dilemma* game. Here, we posit the payoffs as material payoffs and model the true game after individuals' motivations are introduced. Representation of the 2×2 social dilemmas in the previous chapter (Figure 4.4) has the advantage of utilizing the concepts of the normalized fear (F_n), normalized greed (G_n), and normalized cooperators' gain (C_n) in the analyses. The current representation normalizes the absolute cooperators' gain to be 1. The absolute fear and greed are $a - 1$ and b , respectively, and the normalized fear and greed can be calculated by dividing the absolute fear and greed by the payoff range $a - b$: the normalized greed is $\frac{a-1}{a-b}$ and the normalized fear is $\frac{b}{a-b}$.

Players' types are modeled using only one type parameter θ^i , which represents i 's aversion to the outcome in which he defects while the other player cooperates. Figure 5.2 shows the stage game von Neumann-Morgenstern utilities for individuals i and j .

The purpose of restricting θ^i less than $a - b$ is to achieve the maximum comparability with the inequity aversion model. By restricting θ^i less than $a - b$, we preclude a preference-ordering type in which the outcome (D, C) is ranked below the outcome (C, D) . Therefore, given the range of θ^i , an individual can have either a *PD* or an *Assurance* preference type. The game is of incomplete information; θ^i is individual i 's private information. However, the players have a common and objective prior regarding the distribution of types

within the population from which they are randomly and independently drawn. In other words, both players know the cumulative distribution function of types, $F(\theta^i)$, and each player knows that the other player knows the distribution, and each knows that the other player knows that he knows, and so on.

In the finitely repeated 2×2 social dilemma game with uncertainty, the stage game is repeated $N + 1$ times. Each stage is indexed by the number of stages remaining excluding the current one. The game proceeds from stage N to stage 0. The following subsections analyze the equilibria of the game. The focus is whether or not, and under what parameter conditions, the cooperative equilibria of the finitely repeated 2×2 social dilemma game exist.

5.3 Cooperative Equilibria of the Finitely Repeated Game with Two Types of Players

Consider a 2×2 social dilemma game with only two types: the *PD* type for whom $\theta^i = 0$ and the *Assurance* type for whom $\theta^i = \theta$ ($a - 1 < \theta < a - b$). The substantive meaning of the assumption that there exist only two types is that θ^i takes only two values: 0 and θ .² The proportion of the *Assurance* types is δ , which is common knowledge. Our ultimate goal is to examine the existence of, and the conditions for, cooperative equilibria. Before that, we provide a series of definitions and basic propositions.

Definition 22 (Cooperative equilibrium of the stage game) *is a Bayesian equilib-*

²Since the game model is the same as that in Kreps et al., we are carrying on the equilibrium analysis deferred to the readers by Kreps et al. At the same time, this practice will guide us in the analysis of the game when a continuum of types is assumed.

rium in which the Assurance type cooperates and the PD type defects.

Proposition 23 The stage game cooperative equilibrium exists if $\delta \geq \frac{-b}{1-(a-\theta)-b}$.

Proof. When the 2×2 social dilemma game is played only once, the PD type defects since *Defection* is the dominant strategy. Therefore, the focus is under what conditions do the Assurance type players cooperate knowing that all the PD-type players will defect. In the stage game cooperative equilibrium, the expected utility of *Cooperation* for the Assurance type is $\delta \times 1 + (1 - \delta)b$. And, the expected utility of *Defection* is $\delta \times (a - \theta) + (1 - \delta)0$. Therefore, it is rational for an Assurance-type player to cooperate if and only if

$$\delta \times 1 + (1 - \delta)b \geq \delta \times (a - \theta) + (1 - \delta)0 \quad (5.1)$$

$$\delta \geq \frac{-b}{1 - (a - \theta) - b}. \quad (5.2)$$

■

How large a proportion of Assurance types is necessary to sustain a cooperative stage game equilibrium? In other words, what is the smallest value of the right-hand side of (5.2)? The right-hand side of (5.2) increases in $-b$ (absolute fear) and $a - \theta$ (the magnitude of the Assurance-type players' aversion to the outcome $T - S$). Substantively, the stage game cooperative equilibrium requires a higher proportion of the Assurance types as the fear ($-b$) or the value of the outcome (T, S) to the Assurance types ($a - \theta$) increases. For example, set $b = -0.1$, $a = 1.1$, and $\theta = 1.0$. Then, by substituting the specified values of

b , a , and θ to (5.2) we have

$$\delta \geq \frac{0.1}{1 - 0.1 + 0.1} \quad (5.3)$$

$$\delta \geq 0.1. \quad (5.4)$$

This is a case in which *Assurance* types can sustain a stage game cooperative equilibrium even when they constitute only 10% of the population. The reason is that, in this example, the cooperators' gain is relatively large compared to the fear and greed, and the *Assurance* types' orientation toward the mutually cooperative outcome (R, R) is quite strong.

Definition 24 *A cooperative equilibrium of the finitely repeated 2×2 social dilemma game is a sequential equilibrium in which the Assurance type uses the grim trigger strategy and the PD type cooperates until near the end of the game.*

Proposition 25 *If there exists the stage game cooperative equilibrium for a single-stage 2×2 social dilemma game, there always exists a cooperative equilibrium of the finitely repeated 2×2 social dilemma game. In the cooperative equilibrium, the Assurance type uses the grim trigger strategy. If $\delta \geq \frac{a-1}{a}$, the PD type cooperates until stage $n^* + 1$ and defects from stage n^* . If $\delta < \frac{a-1}{a}$, the PD type cooperates until stage $n^* + 2$, uses a mixed strategy at stage $n^* + 1$, and defects from stage n^* , where n^* is the largest integer smaller than $\frac{-b+b\delta}{\delta}$.*

Proof. The key in this proof is showing that for an *Assurance*-type player the grim trigger strategy is rational in the finitely repeated game, provided that there exists the stage game cooperative equilibrium and all other *Assurance* types also use the grim trigger

strategy. In particular, it needs to be shown that it is sequentially rational for *Assurance* types to cooperate at stage n^* in which *Defection* by the *PD* types starts.

Since both types cooperated until the immediately preceding stage (stage $n^* + 1$), the probability that one's partner is of an *Assurance* type still remains at δ . If he cooperates, there is a δ probability that his payoff is 1 in each of the remaining stages including the current one. There is also a $1 - \delta$ probability that his partner is a *PD* type. In that case, he receives a payoff of b in the current stage and 0 in each of the remaining stages. Therefore, the expected utility of *Cooperation* at stage n^* for an *Assurance* type is

$$\delta \times (n^* + 1) + (1 - \delta) \times b. \quad (5.5)$$

On the other hand, if he defects, there is a δ probability that he receives $a - \theta$ in the current stage and 0 for each of the remaining stages. There is also $1 - \delta$ probability that his partner is a *PD* type and his payoff is 0 in each of the stages including the current one. Therefore, the expected payoff of *Defection* at stage n^* for the assurance type is

$$\delta \times (a - \theta) + (1 - \delta) \times 0. \quad (5.6)$$

Cooperation at stage n^* is rational if (5.5) is greater than or equal to (5.6), or

$$\begin{aligned} \delta \times (n^* + 1) + (1 - \delta)b &\geq \delta \times (a - \theta) + (1 - \delta)0 \\ \delta &\geq \frac{-b}{n^* + 1 - (a - \theta) - b}. \end{aligned} \quad (5.7)$$

The existence of the stage game cooperative equilibrium implies that (5.7) is true when $n^* = 0$. Therefore, (5.7) is always true. The possibility that *PD* types may start *Defection* quite early in the game does not give any incentive to an *Assurance*-type player to unilaterally defect in any stage.

Next, we need to verify the sequential rationality of the *PD*-type players in this equilibrium. Since all *PD*-type players are in exactly the same position, the *Defection* point has to be the same for them. Sequential rationality of the equilibrium for the *PD* types implies that (1) it is rational to defect at stage n^* and (2) it is not rational to defect at stage $n^* + 1$. At stage n^* , the expected utility of *Defection* for a *PD* type is $\delta(a) + (1 - \delta)(0) = \delta(a)$. The expected utility of *Cooperation* is $\delta(n + a) + (1 - \delta)b$. Therefore, *Defection* at stage n^* is rational if

$$\begin{aligned} \delta(n + a) + (1 - \delta)b &\leq \delta(a) \\ n^* &\leq \frac{-b + b\delta}{\delta} \\ n^* &\leq \frac{-b}{\delta} + b. \end{aligned} \tag{5.8}$$

On the other hand, starting *Defection* one stage earlier, at stage $n^* + 1$, should not be rational for a *PD* type. The expected utility for the remainder of the game for a *PD* type when she cooperates at stage $n^* + 1$ is $1 + \delta a$. The expected utility for the remainder of the game for a *PD* type when he defects at stage $n^* + 1$ is simply a . *Cooperation* at stage

$n^* + 1$ is sequentially rational for a *PD* type if and only if

$$\begin{aligned} 1 + \delta a &\geq a \\ \delta &\geq \frac{a-1}{a}. \end{aligned} \tag{5.9}$$

In case $\delta < \frac{a-1}{a}$, it is not sequentially rational for a *PD* type to cooperate at stage $n^* + 1$ if all other *PD* types cooperate. But (5.8) shows that it is also not rational to defect at stage $n^* + 1$ for a *PD* type if all other *PD* types also defect. Therefore, the *PD* types use a mixed strategy that makes each *PD* type indifferent between *Cooperation* and *Defection* at stage $n^* + 1$. ■

For example, a set of parameter conditions [$a = 1.2, b = -0.3, \theta = 0.6, \delta = 0.5, n^* = 0$] guarantees the existence of a cooperative equilibrium for both the stage game and the finitely repeated game. In the example, $n^* = 0$ implies that *PD* types defect only in the final stage. When the cooperative equilibrium of a single-stage game exists, the incentive for assurance types to use the grim trigger strategy in the finitely repeated game is even bigger than the incentive to cooperate in the single-stage game.

Proposition 26 *If there does not exist a stage game cooperative equilibrium, the cooperative equilibrium of the finitely repeated 2×2 social dilemma game exists if and only if there exists n^* such that*

$$\frac{-b}{n^* + 1 - (a - \theta) - b} \leq \delta < \frac{-b}{n^* - b}. \tag{5.10}$$

Proof. Because the stage game cooperative equilibrium does not exist, the condition (5.7) is no longer automatically met – there should be n^* that satisfies the condition.

Since $n^* = 0$ contradicts the nonexistence of the stage game cooperative equilibrium, n^* should be greater than or equal to 1. At stage n^* it should be sequentially rational for a *PD* type to defect when all other *PD* types defect. The condition can be given by re-expressing (5.8) in terms of δ :

$$\delta < \frac{-b}{n^* - b}. \quad (5.11)$$

Combining (5.7) and (5.11), we have

$$\frac{-b}{n^* + 1 - (a - \theta) - b} \leq \delta < \frac{-b}{n^* - b}. \quad (5.12)$$

The strategy of the *PD* types is conditional on the relationship between δ and $\frac{a-1}{a}$ in the same manner as that in the proof of Proposition 25. ■

A set of parameter conditions [$a = 1.2, b = -0.5, \theta = 0.4, \delta = 0.3, n^* = 1$] is an example in which the cooperative sequential equilibrium for the finitely repeated game exists in spite that the stage game cooperative equilibrium does not exist. The realized outcome of the cooperative equilibrium depends on the types of the two players. If both players happen to be *Assurance* types, [mutual *Cooperation*] persists to the final stage. The fact that the cooperative equilibrium of the finitely repeated game can exist in the absence of the stage game cooperative equilibrium indicates that even a finite repetition can create incentives for players to cooperate.

When there does not exist the stage game cooperative equilibrium, defection by the *PD* type has to start before the final stage to support the cooperative equilibrium of the finitely repeated game. Otherwise, the final stage is played exactly the same as a single-

stage 2×2 social dilemma game in which no cooperative equilibrium exists. Therefore, both *Assurance* and *PD* types defect and, by backward induction, players defect in every stage. Notice that in (5.10), $n^* = 0$ contradicts the non-existence of the stage game cooperative equilibrium. Therefore, (5.10) rules out the possibility that there exists a cooperative sequential equilibrium of the finitely repeated game in which the *PD* types defect only in the final stage.

5.3.1 Cooperative Equilibrium of the Finitely Repeated Game with a General Distribution of Types

Modeling of a general distribution of types can be done by individualizing θ^i . Each individual has his or her own degree of aversion to exploiting one's partner. In terms of preference ordering, there still exist only two types – *Assurance* and *PD*. But within each preference ordering type, the magnitude of the aversion parameter, θ^i , differs across individuals. A population Θ can be characterized by a cumulative distribution function of types, $F(\theta^i)$. The proportion of the *PD* types is $\delta = F_\theta(a - 1)$ and the proportion of the *Assurance* types is $1 - \delta = 1 - F_\theta(a - 1)$.

The cooperative equilibrium of the finitely repeated 2×2 social dilemma game with a general distribution of types specifies a mapping of each type into the index of the stage in which the type starts defection:

$$Eq : \theta^i \rightarrow n^i. \quad (5.13)$$

For example, $n^i = 1$ implies that a θ^i type player defects at stage 1 and $n^i = 0$ implies that

the type never defects unilaterally unless the other player defected in any of the previous stages.

Let us define p_n to be the proportion of types, in the whole population, who start defection at stage n . Also define q ($q \leq \delta$) to be the proportion of types who cooperate to the final stage unless the other player defects.³ Then,

$$\sum_{n=0}^N p_n = 1 - q. \quad (5.14)$$

An example is provided below, in which there are two defection points: 1 and 0.

Example 27 *A cooperative equilibrium of the finitely repeated 2×2 social dilemma game with a general distribution of types and multiple defection points.*

In this equilibrium, the whole population Θ divides into three subsets: a subset of types who defect at stage 1 (Θ^1), a subset of types who defect at stage 0 (Θ^0), and a subset of types who cooperate to the final stage unless the other player has defected in any of the preceding stages (Θ^C). We can denote the respective proportion of each subset in the whole population as

$$p_1 = \Pr(i \in \Theta^1 | i \in \Theta), \quad (5.15)$$

$$p_0 = \Pr(i \in \Theta^0 | i \in \Theta), \text{ and} \quad (5.16)$$

$$q = \Pr(i \in \Theta^C | i \in \Theta) = 1 - p_1 - p_0. \quad (5.17)$$

³ q is smaller than or equal to δ because being an *Assurance* type is a necessary condition to cooperate in the final stage.

In the cooperative equilibrium,

- All types cooperate from stage N to stage 2.
- $i \in \Theta^1$ starts *Defection* at stage 1. For all $i \in \Theta^1$

$$\frac{-1 + p_1 a}{p_1} < \theta^i \leq \frac{p_1(1-b) - 1}{p_0} + a. \quad (5.18)$$

- $i \in \Theta^0$ starts *Defection* at stage 0. For all $i \in \Theta^0$

$$\frac{p_1(1-b) - 1}{p_0} + a < \theta^i \leq a - 1 - \frac{p_0}{q} b. \quad (5.19)$$

- $i \in \Theta^C$ cooperates to the final stage unless the other player has defected in any of the preceding stages. For all $i \in \Theta^C$,

$$a - 1 - \frac{p_0}{q} b < \theta^i. \quad (5.20)$$

Proof. Let us start with the examination of the sequential rationality condition for a *PD* type that starts *Defection* at stage 1 ($i \in \Theta^1$). For him, *Cooperation* at stage 2 and *Defection* at stage 1 should be sequentially rational. If he defects at stage 2, he will receive a payoff of $a - \theta^i$ at that stage, but his payoff for the remainder of the game is 0.

$$Eu^i(D) = a - \theta^i. \quad (5.21)$$

If he cooperates at stage 2, he is guaranteed to receive a payoff of 1 for that stage.

In the next stage he defects with probability 1. Then there is a p_1 probability that his partner also defects at that stage and his payoff is 0 for that stage. There is also a $1 - p_1$ probability that his partner cooperates at stage 1, thus his payoff at stage 1 is $a - \theta^i$. His payoff for the final stage is sure to be 0 since he defects at stage 1. The expected payoff of *Cooperation* at stage 2 is the respective payoffs combined with the probabilities.

$$Eu^i(C) = 1 + (1 - p_1)(a - \theta^i). \quad (5.22)$$

Cooperation at stage 2 is sequentially rational if and only if (5.21) is greater than or equal to (5.22), or

$$\begin{aligned} 1 + (1 - p_1)(a - \theta^i) &\geq a - \theta^i \\ \theta^i &\geq \frac{-1 + p_1 a}{p_1}. \end{aligned} \quad (5.23)$$

Inequality (5.23) should hold for every type. For example, if $\frac{-1 + p_1 a}{p_1} < 0$ ($\Rightarrow a < \frac{1}{p_1}$) it will hold for every type since the lower limit of θ is 0. To put it differently, if a , the temptation payoff, is too large, it is possible that there exists no cooperative equilibrium of the sort currently being analyzed.

On the other hand, *Defection* at stage 1 should be sequentially rational for him. If he defects at stage 1, the expected utility is

$$Eu^i(D) = (1 - p_1)(a - \theta^i). \quad (5.24)$$

If he cooperates at stage 1, there is a p_1 probability that he will be exploited at that stage and $1 - p_1$ probability that he can enjoy [mutual *Cooperation*] for the current stage. If his partner also cooperates at stage 1, there is a $\frac{p_0}{1-p_1}$ probability that his partner will defect in the final stage and $\frac{q}{1-p_1}$ probability that his partner will cooperate in the final stage. Combining the payoffs and respective probabilities, we have the expected payoff of *Cooperation* for him at stage 1:

$$Eu^i(C) = p_1(b) + (1 - p_1) \left[1 + \frac{p_0}{1 - p_1}(0) + \frac{q}{1 - p_1}(a - \theta^i) \right]. \quad (5.25)$$

It is rational to defect at stage 1, if (5.24) is greater than or equal to (5.25), or

$$\begin{aligned} (1 - p_1)(a - \theta^i) &\geq p_1(b) + (1 - p_1) \left[1 + \frac{p_0}{1 - p_1}(0) + \frac{q}{1 - p_1}(a - \theta^i) \right] \\ \theta &\leq \frac{p_1(1 - b) - 1}{p_0} + a. \end{aligned} \quad (5.26)$$

Next, for those who cooperate at stage 1, but defect at stage 0, *Cooperation* should be sequentially rational at stage 1 and *Defection* should be sequentially rational at stage 0. The first condition is the reverse of (5.26).

$$\theta > \frac{p_1(1 - b) - 1}{p_0} + a. \quad (5.27)$$

The second condition can be calculated as follows. In the final stage, when the partner has cooperated in the previous stage, the expected utility of *Cooperation* is

$$\frac{q}{p_0 + q}(1) + \frac{p_0}{p_0 + q}b. \quad (5.28)$$

and the expected utility of *Defection* is

$$\frac{q}{p_0 + q}(a - \theta^i) + \frac{p_0}{p_0 + q}(0). \quad (5.29)$$

For *Defection* to be sequentially rational, (5.28) has to be greater than or equal to (5.29),

or

$$\begin{aligned} \frac{q}{p_0 + q}(a - \theta^i) &\geq \frac{q}{p_0 + q}(1) + \frac{p_0}{p_0 + q}b \\ \theta^i &\leq a - 1 - \frac{p_0}{q}b. \end{aligned} \quad (5.30)$$

Finally, the reverse of (5.30) is the condition for i to cooperate in the final stage if the partner has cooperated in the previous stage.⁴

$$\theta^i > a - 1 - \frac{p_0}{q}b. \quad (5.31)$$

■

The existence of multiple defection points across the *PD* types increases the incentive for an *Assurance* type player to use the grim trigger strategy. In the example above, an *Assurance*-type player who has experienced [mutual *Cooperation*] in stage 1 knows that the probability of his partner now being an *Assurance* type is $\frac{\delta}{1-p_1}$, which is greater than δ . Therefore, even when *Cooperation* is not a rational strategy for him in a single-stage

⁴Since $\theta^i > a - 1$ is the condition for i to be an *Assurance* and the threshold value of $\theta^i = a - 1 - \frac{p_0}{q}b$ is greater than $a - 1$, being an *Assurance* type does not guarantee that a player will cooperate in the final stage even when his partner has cooperated in the previous stage. In other words, being an *Assurance* type is a necessary but not sufficient condition to cooperate to the final stage. On the other hand, we can also see that the threshold is lower than that for being a cooperator in the stage game cooperative equilibrium.

game, it can be rational in the final stage of a finitely repeated game.

5.3.2 Hybrid Equilibria of the Finitely Repeated Game

In between [all *Defection*] equilibrium and the cooperative equilibrium, there also exist a series of hybrid equilibria.

Definition 28 *A hybrid equilibrium of the finitely repeated 2×2 social dilemma game is a sequential equilibrium in which all types start the game with Defection and, at a later stage, all types switch to the cooperative equilibrium for the remaining subgame.*

Suppose that a single-stage 2×2 social dilemma game is repeated $N + 1$ times and there exists the cooperative equilibrium for the finitely repeated game. Suppose further that all types have defected from the first stage to stage m . In that case, the remaining subgame is an m times repeated game of the stage 2×2 social dilemma game. A transition from [mutual *Defection*] to [mutual *Cooperation*] is possible in equilibrium if m is greater than $\max(n^i)$, the first stage in which *Defection* by any type occurs in the cooperative equilibrium for the entire game.

The intuition behind this equilibrium is as follows. When all the types defect from the first stage, there can be no belief update regarding the type of one's partner. Then, at a later stage, the initial prior that supports the existence of the cooperative equilibrium is intact and may still support a cooperative sequential equilibrium for the remaining subgame. Therefore, the two players can play the remaining subgame as if it were a new game of finite repetition.

In real settings, the transition may not be coordinated by the two players without communication. Therefore, it will probably involve some stages in which at least one player's behavior is not optimal. However, after playing [mutual *Defection*] for a number of stages, all types have an incentive to send a signal to the other player, by cooperating in the current stage, that they want to coordinate on [mutual *Cooperation*]. The signaling is worth trying when there are enough number of stages remaining to compensate the possible loss.

On the part of the players who receive the signal, whether or not the signal sender is an *Assurance* type matters only to an extent. Insofar as there are enough number of stages remaining and it appears clear that the signal sender wants to move to [mutual *Cooperation*], there is a good opportunity to reap the gains from [mutual *Cooperation*] for a sustained time period.

5.4 Conclusion

This chapter analyzed the equilibria of the finitely repeated 2×2 social dilemma game when there are multiple types of players. A player's type is defined by the way he or she transforms the material payoffs of an action situation into von Neumann–Morgenstern utilities of a game. The equilibrium analyses are conducted first, assuming that there are only two types and second, assuming that there is a continuum of types. In both cases, there does exist the cooperative equilibrium of the finitely repeated game whenever there exists the stage game cooperative equilibrium. When there does not exist the stage game cooperative equilibrium, the possibility of cooperation in the finitely repeated game depends on the *Defection* point of the *PD*-type players when the *Assurance* types use the grim

trigger strategy, and there being multiple *Defection* points among the *PD* types that facilitate the *Assurance* types' use of the grim trigger strategy. There also exist a series of hybrid equilibria in which a transition from mutual *Defection* to mutual *Cooperation* occurs. Though the exact replication of this kind of equilibria in real settings is not very likely, it still provides a rational basis for the risky investment/initiation by the players who want to escape from the trap of mutual *Defection*. As was the case in the analyses of the static models in Chapter 4, the possibility of cooperation is affected by the environment of an action situation – the material payoff structure of a 2×2 social dilemma – as well as the culture of a group – distribution of types within the population. In addition, with the existence of multiple equilibria, the problem of coordination becomes more significant.

Chapter 6

Heterogeneity and Interdependence

6.1 Introduction

This chapter tests the equilibrium analyses of Chapter 5 using an experimental data set of finitely repeated 2×2 social dilemmas. Chapter 4 has shown that there is a significant proportion of individuals who are not entirely selfish. The nonselfish individuals are most likely to have *Assurance*-type preferences of which the behavioral principle is reciprocity. Chapter 5 has shown that when there are multiple types of players in a finitely repeated 2×2 social dilemma game, there are equilibria other than [*all Defection*] even when there does not exist the stage game cooperative equilibrium.

The empirical tests of this chapter focus on the impacts of the theoretical variables analyzed in Chapter 5 on the strategy choice of players at each stage of the finitely repeated

2×2 social dilemma games. At any given stage, a player's choice is affected by several factors.

One's own type θ^i . So far, we have used the term "type" in two contexts. In the first context, *type* refers to the specific value of a player's type parameter in his utility function that transforms allocational states into utilities. In the second context, *type* in the "preference type" is defined by the characteristics of a player's preference ordering over the possible outcomes of an action situation. *Type* in the first sense, in conjunction with the structure of the material payoffs, determines *type* in the second sense. Unlike other theoretical variables discussed in Chapter 5, a player's type is not directly observable; it has to be inferred from the actions taken during the game.

The material payoff structure of a stage, $\pi = (F_n, G_n, C_n)$, is another factor that affects a player's choice at the stage. In Chapter 4, we have used the concepts of *Fear*, *Greed*, and *Cooperators' Gain* to characterize material payoff structures. The use of the representation, $\pi = (a, b)$, was necessary in Chapter 5 to simplify the repeated game equilibrium analyses. In this chapter, we return to the concepts of *Fear*, *Greed*, and *Cooperators' Gain* because of their general applicability. In the data to be analyzed, the material payoff structure varies across stages, adding another complication to the repeated game. However, the basic conclusions of the equilibrium analyses in Chapter 5 still hold; i.e., there exists the cooperative equilibrium of the finitely repeated 2×2 social dilemma game, and the behavior of a player in the equilibrium is a function of the number of stages left, his own type, and the material payoff structure.

The stage index, t , matters in two ways. First, in the cooperative equilibrium,

the possibility of *Defection* by either of a paired two players increases as the repeated game approaches the final stage. On the other hand, when the two players are trapped in [mutual *Defection*], the initiation of *Cooperation* by any of the two players is easier when there are more stages left. This is because when both of the players have played *Defection* from the beginning to the current stage, there is no belief update and the remaining stages can be played as a repeated game with a smaller number of stages than the original one. The coordination problem suggests that the transition to the cooperative equilibrium for the remaining stages will not be easy. On the other hand, when the players learn the futility of [mutual *Defection*] and take the chance of initiating *Cooperation*, the transition is plausible. The more stages that are left, the easier will be the transition.

The history of the game before the stage t , $h(t)$. In the cooperative equilibrium the way a player has played in the previous stages conveys information about his type to the partner. The players are assumed to have a common prior, $F(\theta^i)$. A player's updated belief at any stage of the game regarding his partner's type can be the same as the original prior, as is the case when the two players are trapped in the mutual *Defection* or in the earlier phases of the cooperative equilibrium in which both players cooperate regardless of their types. On the other hand, in the stage of the cooperative equilibrium in which behavior of different types diverge, a player's action changes the partner's belief about his type.

The way the immediately preceding stage is played also matters when the transition from mutual *Defection* to mutual *Cooperation* is attempted by the players. Chapter 5 discussed three types of equilibria: (1) the cooperative equilibrium in which two players start the game with *Cooperation* and continue to cooperate until near the final stage when

the behavior depends on one's type, (2) the all *Defection* equilibrium in which both players defect from the first to the final stage, and (3) the hybrid equilibria in which two players start the game with *Defection* but convert to mutual *Cooperation*.

In reality, it is not very likely that two players convert to *Cooperation* at exactly the same stage. Therefore, the transition may involve at least one stage of non-equilibrium play. A player's action during the transition phase can convey his/her intention to his partner regarding equilibrium choice. For example, a player unilaterally cooperating after mutual *Defection* has been played for a while may intend to send a message that says "Shall we move to the cooperative equilibrium?" If in the next stage the other player reciprocates the initiation by cooperating, she is sending a message that "I would love to and thanks for the initiation." If a player persists on *Defection* even when the other player has sent the cooperative message, he might be saying "Even though you have shown me the intention of moving to the cooperative equilibrium, I am not sure about your type and I would rather stay in this mutual *Defection* than taking the risk of being exploited later." A complete specification of the ways in which all the previous stages are played out is not plausible in regression analyses. Instead, we will use s^{t-1} , the outcome of the immediately preceding stage.

In general, we can model the probability of a player i cooperating at stage t as

$$P(s^{i,t} = C) = f(\pi^t, s^{t-1}, t : \theta^i) \quad (6.1)$$

where ω^t is characterized by G_n, F_n , and C_n , $s^{t-1} = (s^{i,t-1}, s^{j,t-1})$ is the way the game is played out in the immediately preceding stage, t is the index for the current stage, and θ^i

is player i 's type.

6.2 Experimental Design and Procedure

The experiment was conducted during the fall semester of 1998 at Indiana University. This chapter analyzes a subset of the data generated by the experiment. Schmidt et al. (forthcoming) report the design and analyses of the entire experiment. The subjects of the experiment were paid volunteers recruited from an undergraduate microeconomics class. Prior to volunteering, subjects were informed that they would participate in a decision-making experiment in which they would be paid, in cash, an amount dependent upon their decisions and the decisions of others in the experiment. Subjects were recruited by cohorts of eight in each experimental session. Upon arriving at the laboratory, subjects were randomly seated at computer monitors. Each subject in an experimental session was anonymously assigned a one-digit identification number. Subjects knew their own identification number, but could not associate any other subject in the room with that subject's identification number.

After receiving an identification number, the subjects were informed of the conditions for the experiment. First, they read through a set of computerized instructions regarding the structure of the experiment. The experimenter then presented further instructions explaining the decision-making environment publicly on an overhead projector. The subjects did not know at any time, before or after the experiment, with whom they were matched. They were also told that their decisions would remain anonymous during and after the experiment.

The games played were described to the subjects as board games with a row and column player. Each subject made a choice between option "C" – labeled as "B" during the experiment – and option "D" – labeled as "A" during the experiment. Each subject always saw himself/herself as a row player in the computer monitor. (Reference of a player as *Row* or *Column* later during the analyses of the experimental data is based on the way a subject is shown to the experimenters on their monitoring computer screen.)

Subjects were informed that at the end of the experimental session they would privately receive their earnings from the experiment plus a \$5 show-up fee. They were instructed to think of the payoffs in the individual stages as "computer pesos," where the conversion rate was 100 pesos equal \$1.

The experiment was divided into two phases: Phase 1 and Phase 2. In each of the two phases, each subject was randomly matched with another person at the beginning and played a repeated game with 12 stages with the same subject he or she was matched with at the beginning of the session.¹ After Phase 1, subjects were divided into two groups of four based on the number of cooperative choices they made in Phase 1. At the beginning of Phase 2, an overhead was presented to the subjects publicly showing the assignment of each subject to groups and each player and his/her Phase 1 partner's number of *Cooperation* (called B during the experiment) and *Defection* (called A, during the experiment). Then, each player was randomly matched with one of the three other players in his/her group and

¹Usually a two-person repeated game implies that a pair of players play the entire stages of a repeated game without replacing the partner. With the introduction of an experimental design in which a group of more than two players plays a certain number of stage games with different partners at each stage, the original repeated game in which no replacement of partner occurs is called fixed matching repeated game, while the other is called random matching multiple-stage game. The experimental data analyzed in this chapter is that of the fixed matching repeated game.

		Player 2	
		D	C
Player 1	D	40 40	110 10
	C	10 110	80 80

Game 1

		Player 2	
		D	C
Player 1	D	50 50	110 10
	C	10 110	70 70

Game 2

		Player 2	
		D	C
Player 1	D	50 50	110 10
	C	10 110	90 90

Game 3

		Player 2	
		D	C
Player 1	D	60 60	110 10
	C	10 110	80 80

Game 4

		Player 2	
		D	C
Player 1	D	30 30	110 10
	C	10 110	70 70

Game 5

		Player 2	
		D	C
Player 1	D	40 40	110 10
	C	10 110	60 60

Game 6

Figure 6.1: Material Payoff Structure of Stage Games (*Source*: Schmidt et al., forthcoming)

played another repeated game with 12 stages without replacing the partner.

In Phase 2, a player can tell whether he/she belongs to the group of players with higher numbers of cooperative choices by looking at the numbers of cooperative choices made by other individuals in the group and comparing them with the numbers of cooperative choices made in Phase 1 by the individuals in the other group. But he/she cannot match subject number to one of the players in the room. The material payoff structures of the stage games are shown in Figure 6.1. In both phases, the games were played in a sequence of 1-2-3-4-5-6-1-2-3-4-5-6. At each stage, two players made their choices simultaneously. Subjects were given complete information about the pecuniary payoff structure, the number of stages in a phase, and the rematching condition in Phase 2.

	<i>High Cooperators' Gain</i>	<i>Low Cooperators' Gain</i>
<i>Greed = Fear</i>	Game 1 ($C_n = 0.4, G_n = 0.3, F_n = 0.3$)	Game 2 ($C_n = 0.2, G_n = 0.4, F_n = 0.4$)
<i>Greed < Fear</i>	Game 3 ($C_n = 0.4, G_n = 0.2, F_n = 0.4$)	Game 4 ($C_n = 0.2, G_n = 0.3, F_n = 0.5$)
<i>Greed > Fear</i>	Game 5 ($C_n = 0.4, G_n = 0.4, F_n = 0.2$)	Game 6 ($C_n = 0.2, G_n = 0.5, F_n = 0.3$)

Table 6.1: Normalized Material Payoff Parameters

6.3 Overall Results

The overall rate of cooperation in the experiment was 32.5%. Figure 6.2 presents the frequency of cooperation across decision stages. In contrast to Andreoni and Miller (1993) and Selten and Stoecker (1986), there is no clear decline of *Cooperation* by the end of the game. This may be due to the differences between the current and the aforementioned two experiments. The two studies used a single material payoff matrix for the entire experiments of which the *Cooperators' Gain* was quite small. The equilibrium analyses of Chapter 5 suggest that the frequency of *Cooperation* in a finitely repeated 2×2 social dilemma game depends heavily on the material payoff structure. Another possible explanation is that while the aforementioned two experiments utilized a long series (20) of supergame of which one game is a 10-stage, finitely repeated 2×2 social dilemma game, the current experiment uses only a single repeated game in each of the two phases.

The regrouping of the subjects in Phase 2 into the two groups of high and low cooperators may have opposing impacts on their behavior in Phase 1. Though the experimental instruction states it in a neutral way – regrouping for Phase 2 will be based on your choices of A and B in Phase 1 –, subjects could conjecture that by cooperating more in Phase 1, they have a better chance of belonging to the high cooperators' group in Phase 2.

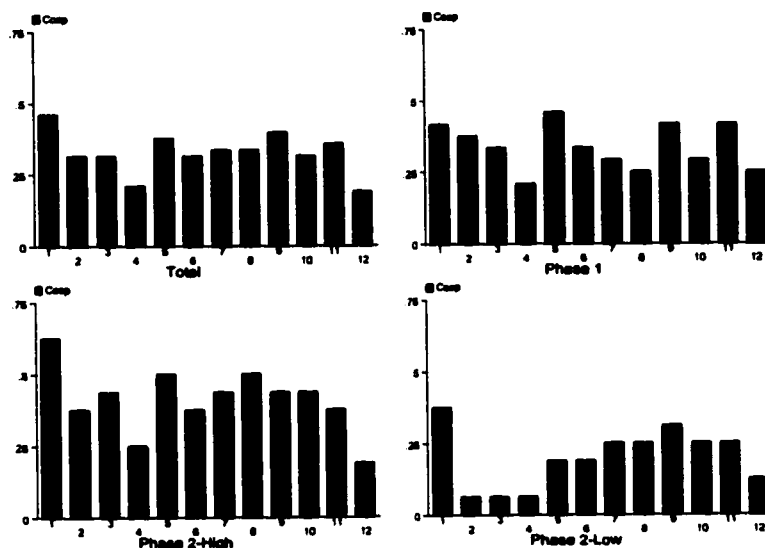


Figure 6.2: Frequency of *Cooperation* across Stages

In that sense, players have an incentive to cooperate more in Phase 1 with hopes of being matched with an *Assurance* type in Phase 2. The flip side of the incentive is that when a player defects, his/her partner may not immediately retaliate in the next stage, thus making *Defection* in Phase 1 less costly.

However, the evidence suggests that the strategic consideration for Phase 2 group belonging is not as significant as player type in determining subjects' behavior in Phase 1. If the strategic consideration was a primary factor affecting Phase 1 behavior, we would have to see no significant difference in the frequency of *Cooperation* between the high and low cooperators' group in Phase 2. But the average *Cooperation* rate is more than twice higher in the high cooperators group (41%) than in the low cooperators', indicating that the behavior of individuals in Phase 1 is more likely to reflect their true types than their strategic considerations for Phase 2 group belonging.

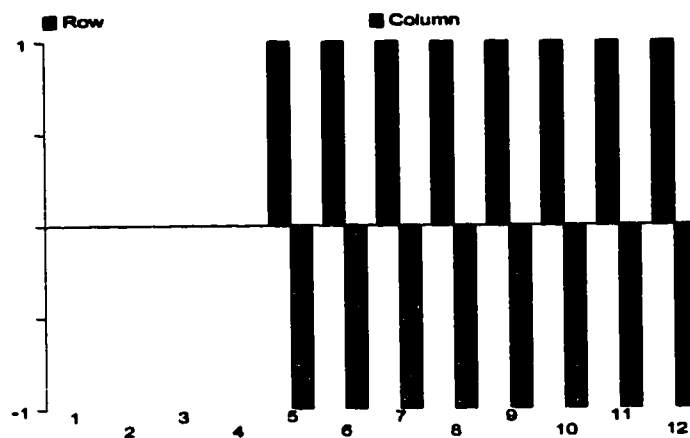


Figure 6.3: Decisions Made by a Pair of Players

Next, we would like to see how well the equilibrium analyses of Chapter 5 describe the actual unfolding of the repeated games. Figure 6.3 shows the way the outcome of a game will be represented heretofore. In the figure, x -axis represents stages indexed from 1 to 12 and y -axis represents players' strategy choices. The presence of a bar extending upward(downward) indicates that the *Row*(*Column*) player has cooperated in that stage. In Figure 6.3, which shows the outcome of the game played by a pair of players in Phase 1 of session 1, both the players defect from stages 1 to 4 and cooperate from stages 5 to 12.

Figures 6.4 to 6.12 show all the decisions made by pairs of players in Phase 1, Phase 2 low cooperators' group, and Phase 2 high cooperators' group. In the figures, each player is identified by a two-digit number of which the first digit is the session number and the second digit is the subject identification number assigned to the player in that session. For example, Player 11 is a subject in Session 1 with subject identification number 1.

Each game is titled with a four-digit number in which the first(second) two digits correspond to the *Row*(*Column*) player's identification number. For example, the title 1112

Figure 6.4: Decisions in Phase 1, Session 1

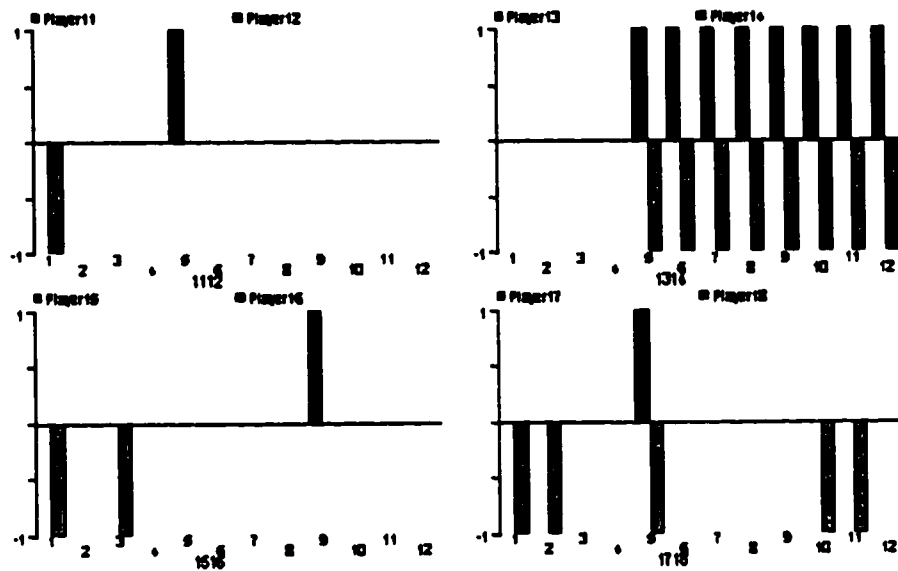


Figure 6.5: Decisions in Phase 1, Session 2

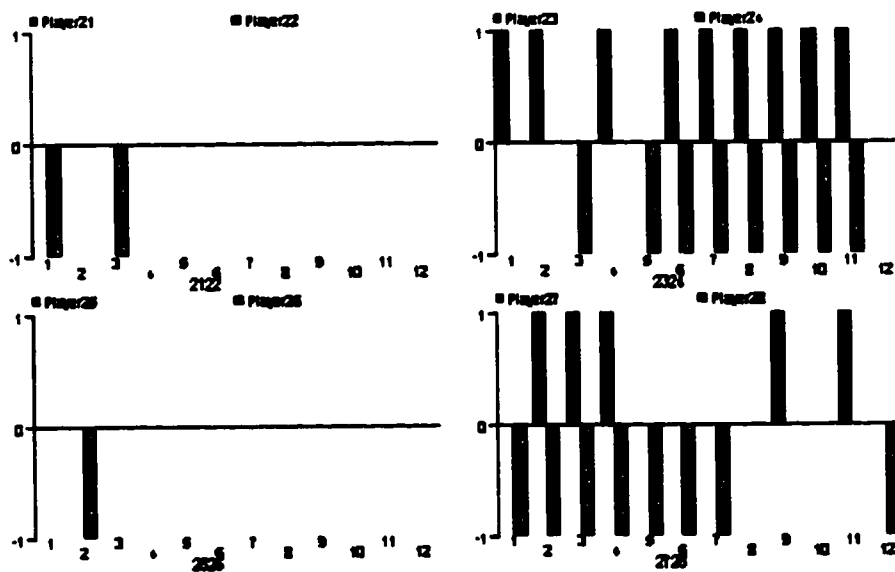


Figure 6.6: Decisions in Phase 1, Session 3

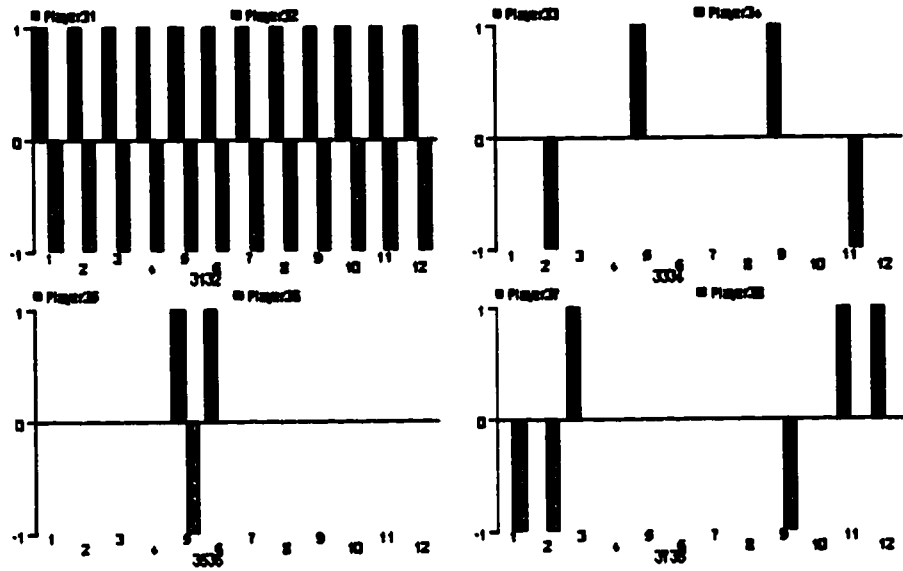


Figure 6.7: Decisions in Phase 2 Low Cooperators' Group, Session 1

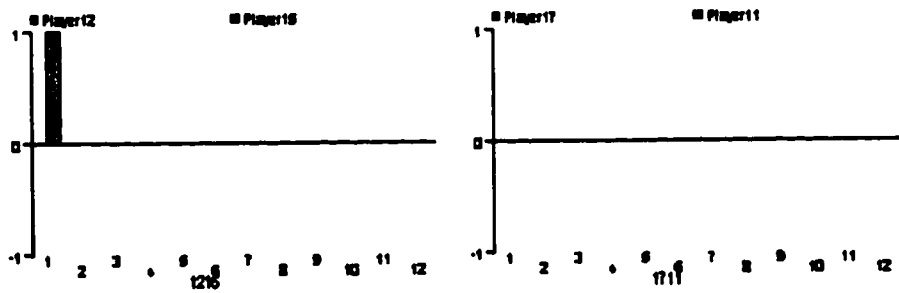


Figure 6.8: Decisions in Phase 1 Low Cooperators' Group, Session 2

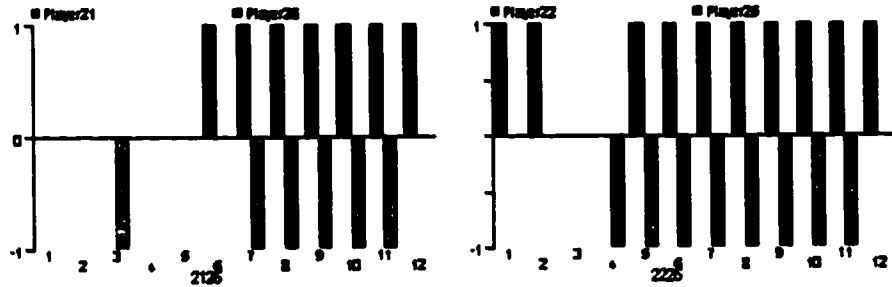


Figure 6.9: Decisions in Phase 2 Low Cooperators' Group, Session 3

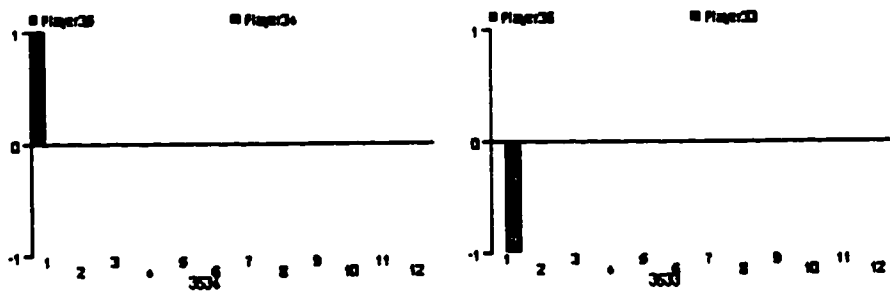


Figure 6.10: Decisions in Phase 2 High Cooperators' Group, Session 1

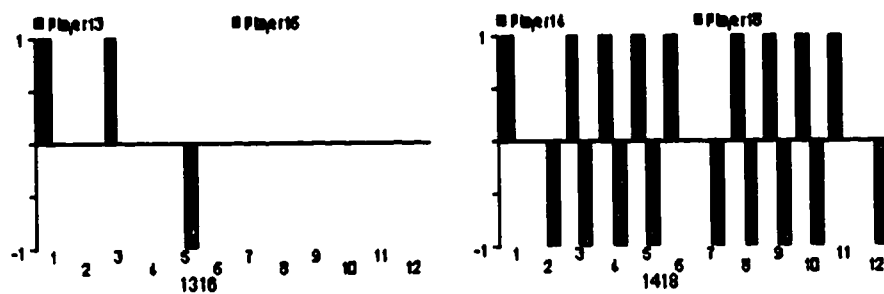


Figure 6.11: Decisions in Phase 2 High Cooperators' Group, Session 2

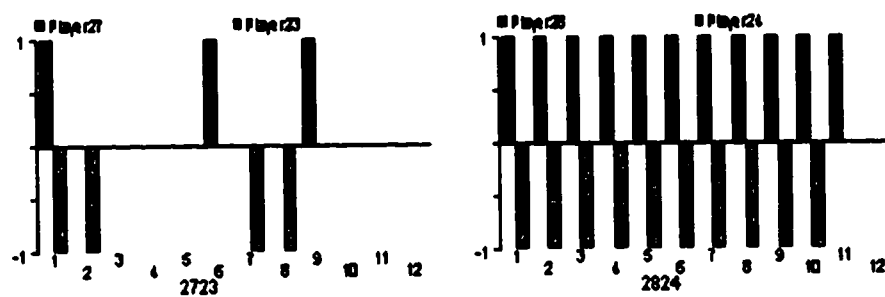
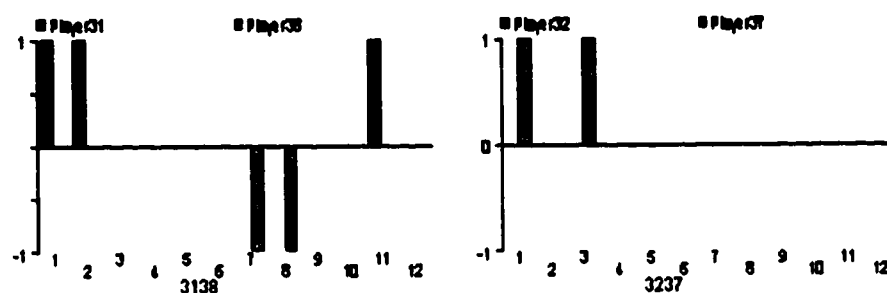


Figure 6.12: Decisions in Phase 2 High Cooperators' Group, Session 3



for the first game of Session 1, Phase 1 shown in Figure 6.4 implies that the game was played by Player 11 and Player 12. The titles for individual games help to identify any of the 24 players' 24 total decisions. For example, Player 11's 12 decisions in Phase 1 are shown in the upper half of the game titled 1112 in Figure 6.4. His/her 12 decisions in Phase 2 are shown in the lower half of the game titled 1711 in Figure 6.7.²

The outcomes – the actual ways the repeated games were played – are divided into four patterns:

1. equilibrium outcomes that meet the strict criteria of the sequential equilibrium;

²Though in each game diagram, *Column* player's choice in each stage is represented slightly to the right of *Row* player's choice, potentially creating an image to the readers that the stage game was conducted sequentially, the actual plays were simultaneous; two paired players made their choices at the same time.

Pairing of Player 11 with Player 12, Player 13 with Player 14, and so on in Phase 1 of each experimental session may raise a question regarding the randomness of the pairing rule in Phase 1. However, the pairing was random in the sense that the identification numbers were randomly assigned to them before the matching, and the subjects knew neither the identification number of the player with whom they were paired nor the personal identity of the partner.

2. outcomes in which the cooperative equilibrium for a subgame is reached through a non-equilibrium path;
3. outcomes in which the [*all Defection*] equilibrium for a subgame is reached through a non-equilibrium path; and
4. noisy outcomes that do not fit into one of the above three patterns.

Strict sequential equilibrium outcomes: Recall the three types of sequential equilibrium discussed in Chapter 5. In the cooperative equilibrium both players start the game with *Cooperation* and cooperate until near the final stage when behavior depends on one's type. In the [*all Defection*] equilibrium, both players defect from the first to the last stage. In a hybrid equilibrium, both the players start the game with *Defection*, but they convert to a cooperative equilibrium at a point in the game when there are enough stages left to go to support the cooperative equilibrium for the subgame.

Of the 24 repeated games, only four have outcomes that meet the strict criteria of the sequential equilibrium. Interestingly, however, the outcomes of the four games include all of the three major patterns of equilibrium. Game 3132 of Phase 1, Session 3 (see Figure 6.6) and Game 2824 of Phase 2 high group, Session 2 (see Figure 6.11) are the cases in which the cooperative sequential equilibrium is played. Two players in Game 3132 and player 28 in Game 2824 are most likely to be *Assurance* types. But player 24 defects in stage 11 when the partner has cooperated in stage 10. Therefore, the outcome of the game 2824 is the cooperative sequential equilibrium played by a *PD* type and an *Assurance* type.

Game 1711 of Phase 2 low group, Session 1 (see Figure 6.7) is the only case in which the strict [*all Defection*] equilibrium is played. Neither of players 11 and 17's decisions in

Phase 1 indicate the possibility that they are *Assurance* type players unfortunately stuck in the [*all Defection*] equilibrium; they are most likely to be *PD* types.

Game 1314 of Phase 1, Session 1 (see Figure 6.4) is a case of hybrid equilibrium played, most likely, by two *Assurance*-type players. Both the players defect from stage 1 to stage 4 and cooperate from stage 5 to the final stage. The players were lucky in that they chose to convert to *Cooperation* at exactly the same stage.

Next, we see the outcomes that do not meet the sequential equilibrium conditions. The focus is whether or not the course of play converges to an equilibrium pattern over time and, if it does, to which equilibrium. At any stage, except at the stage of the cooperative equilibrium in which the behavior of different types diverges, both players choose the same action in a sequential equilibrium. Therefore, any mismatch of choices at earlier stages of a game is a sign that the game is not played consistently with one of the sequential equilibria. However, since there are multiple equilibria, and communication between the players is not allowed, we can expect many cases in which two players start the game with different equilibria in mind. In other words, non-equilibrium outcomes can occur even when the two players of a repeated game are both rational and try to play the game with equilibrium strategies.

A non-equilibrium play of the repeated game in earlier stages implies a stage game outcome of (D, C) or (C, D) , and the preference-ordering types we modeled (based on the evidence discussed in Chapter 4) do not allow unconditional *Cooperation* to be rational. Therefore, whenever there occurs a non-equilibrium stage outcome, we expect a rapid convergence to one of the sequential equilibria defined for a subgame. To which

equilibrium, and how fast the actual behavior of the two players converge, will depend upon the (im)patience of the cooperator and (non)responsiveness of the defector in the stage with the non-equilibrium outcome.

From non-equilibrium play to a cooperative equilibrium: Game 2324 (in Figure 6.5), Game 2126 (in Figure 6.8), Game 2225 (in Figure 6.8). In this pattern of outcomes, the game starts with players' non-equilibrium behavior and reaches the cooperative equilibrium for a subgame. For example, in Game 2225, Row player starts the game with *Cooperation* while Column player starts the game with *Defection*. Row player still cooperates in stage 2, but Column defects. In stage 3, Column reciprocates Row's cooperation in the previous two stages with *Cooperation*, but Row plays *Defection* after having been exploited twice in the previous stages. Then in stage 4, Row again plays *Cooperation* having seen Column's *Cooperation* in stage 3. But Column has already switched back to *Defection*. This mal-coordination stops at stage 6 when both players play *Cooperation*. From that point on, the two players play *Cooperation* to stage 11. The outcome for the subgame that consists of stages 6 to 12, therefore, can be viewed as a result of cooperative equilibrium play. Game 2324 and Game 2126 show similar patterns.

From non-equilibrium play to [all Defection] equilibrium: Game 1112 (in Figure 6.4), Game 2122 (in Figure 6.5), Game 1215 (in Figure 6.7), Game 3534 (in Figure 6.9), Game 3633 (in Figure 6.9), Game 1316 (in Figure 6.10), Game 3237 (in Figure 6.12). In this pattern of outcomes, the mismatch of choice in earlier stage(s) rapidly converges to the [*mutual Defection*] equilibrium for a subgame. The three games in the Phase 2 low cooperators' group are the typical cases. When the last *Cooperation* is preceded by at

least one stage in which both players defected, the incidence of *Cooperation* can also be viewed as a futile trial to make a transition from [*mutual Defection*] equilibrium to the cooperative equilibrium.

Noisy outcomes: There are also many outcomes that do not fit into one of the three patterns discussed above. On a descriptive level, they can be divided into two groups. Those noisy outcomes in Phase 1, in which unilateral cooperative choices occur near and in the last stage, can be viewed as the cases in which one of the players cooperates to increase his or her chance to belong to a high cooperators' group in Phase 2. Those outcomes in the Phase 2 high cooperators' group – in which number of cooperative choices are not small, but there are too many (C, D) or (D, C) outcomes – can be viewed as cases of persistent coordination failure.

6.4 Heterogeneity: Fixed Effects Logit Analyses

Sections 6.4 and 6.5 analyze the experimental data with a series of probit and logit models to establish the causal impacts of the theoretical variables on the players' choice of *Cooperation* and *Defection* at each stage. Two special characteristics of the experimental data do not allow us to conduct the standard procedures of the logit and probit estimations.

First, the data have a panel structure. There are 24 individuals, each of whom makes 24 decisions. Conducting the standard binary choice estimations would imply that the 24 individuals are homogeneous in terms of their propensity to cooperate. The fundamental assumption of this study has been that individuals are heterogeneous in their motivations. Therefore, an individual's choice at a stage is more likely to be correlated to

his or her choices in other stages than it is to the choices made by others.

Second, a player is always paired with another individual for the entire 12 stages of a phase. In equilibrium, any two paired players' choices have to be the same – except in the stage of the cooperative equilibrium, in which different types' behavior diverge. The mismatch between the choices of two rational players can happen in a couple of ways. It can result from a coordination failure, in which one player plays the cooperative equilibrium strategy and the other plays the [all *Defection*] equilibrium strategy. Or, it can happen when a player initiates the transition to the cooperative equilibrium from the [mutual *Defection*]. In both cases, a rapid coordination is expected.

In sum, the 576 observations cannot be treated as independent from each other due to the problems of heterogeneity and interdependence. Unfortunately, currently available estimation procedures for the binary choice data do not allow us to address both of the problems within a single estimation.³ Therefore, we control for the two problems separately. First, in this section, a series of (conditional) fixed effects logit analyses are conducted to control for the inter-individual heterogeneity in the panel data. The problem of interdependence between two paired players' choices will be addressed in the next section with a series of bivariate probit analyses.

Both the fixed effects logit and bivariate probit analyses of this chapter are based on a reduced form random utility model. The random utility model, in the context of binary choice problems, assumes that an individual's choice reflects differences in the expected utilities of the two choice options. In the 2×2 social dilemma game, the difference is that

³There is a statistically less sound way of taking both of the problems into account by including individual dummy variables in the bivariate probit analyses. The model is estimated in an exploratory manner at the end of this chapter.

between the expected utilities of *Cooperation* and that of *Defection*. Drawing on (6.1) in Section 1, the respective utility of *Cooperation* and *Defection* for individual i at stage t can be expressed as

$$u^{i,t}(C) = f_c(\pi^t, s^{t-1}, t; \theta^i)$$

$$u^{i,t}(D) = f_d(\pi^t, s^{t-1}, t; \theta^i)$$

in which π^t is the structure of the pecuniary payoffs at stage t , s^{t-1} is the outcome of the previous stage, t is the stage index, and θ^i is individual i 's type.

The structural estimation of the model requires that the utility functions f_c and f_d be specified according to the motivational models discussed in Chapter 4. However, due to the identification problems, this chapter estimates reduced form linear random utility models in which the expected utilities are specified as

$$u^{i,t}(C) = \alpha_{i,c} + \beta'_c \mathbf{x} + \epsilon_{i,c}$$

$$u^{i,t}(D) = \alpha_{i,d} + \beta'_d \mathbf{x} + \epsilon_{i,d},$$

where \mathbf{x} is a vector including π^t , s^{t-1} , t and other independent variables.

Let $u^{i,t}$ denote $u^{i,t}(C) - u^{i,t}(D)$, the difference between the expected utilities of *Cooperation* and *Defection*. Then,

$$\begin{aligned}
u^{i,t} &= (\alpha_{i,c} + \beta'_c \mathbf{x} + \epsilon_c) - (\alpha_{i,d} + \beta'_d \mathbf{x} + \epsilon_d) \\
&= (\alpha_{i,c} - \alpha_{i,d}) + (\beta'_c - \beta'_d) \mathbf{x} + (\epsilon_c - \epsilon_d) \\
&= \alpha_i + \beta' \mathbf{x} + \epsilon,
\end{aligned} \tag{6.2}$$

in which $\alpha_i = \alpha_{i,c} - \alpha_{i,d}$, $\beta' = \beta'_c - \beta'_d$, and $\epsilon = \epsilon_c - \epsilon_d$.

In the data, what we observe is not $u^{i,t}$, but the binary variable $y^{i,t}$, which takes a value of 1 if individual i cooperates at stage t and 0 otherwise. Therefore, our observation is

$$\begin{aligned}
y^{i,t} &= 1 && \text{if } u^{i,t} > 0 \\
y^{i,t} &= 0 && \text{if } u^{i,t} \leq 0.
\end{aligned}$$

The probability that individual i will cooperate is

$$\begin{aligned}
&\Pr(u^{i,t} > 0) \\
&= \Pr(\alpha_i + \beta' \mathbf{x} + \epsilon > 0) \\
&= \Pr(\epsilon > -\alpha_i - \beta' \mathbf{x}).
\end{aligned} \tag{6.3}$$

We can innocently assume that ϵ has a *standard* logistic or normal distribution because utility is defined up to linear transformation. Then, due to the symmetry of the

normal and logistic distributions, (6.3) can be rewritten as

$$\Pr(\epsilon < \alpha_i + \beta' \mathbf{x}) = F(\alpha_i + \beta' \mathbf{x}). \quad (6.4)$$

In the probit model, F is the standard normal cumulative distribution function. In the logit model, F is the logistic cumulative distribution function.

The individual specific term α_i in (6.2) represents the heterogeneity in the fixed effects model. A literal interpretation of (6.2) would imply that the difference among individuals lies in their baseline propensity to cooperate and the covariates (\mathbf{x}) have identical impacts on individuals' behavior. The theoretical analyses in Chapters 4 and 5, however, suggested that the inter-individual heterogeneity does not exist in a linear fashion. Thus, (6.2) is not exactly correct; individuals may differ fundamentally in terms of the ways they respond to the changes in the covariates \mathbf{x} . In other words, β' may also need to be individualized – an impossible task because, in that case, the number of parameters to be estimated increases in proportion to the number of observations.

A direct implementation of (6.2) in the standard logistic model can be done by including a dummy variable for each individual. However, the maximum number of observations that can be used in estimating each of the coefficients for individual dummy variables is only 24, which is not large enough for the coefficients to be reliable.⁴

This study addresses the problem of heterogeneity with a series of (conditional)

⁴In a linear panel model, the estimates of the coefficients for individual dummy variables and those for the common independent variables are asymptotically independent. Therefore, when the number of observations is large, the latter are not biased even when the former are. However, the binary choice models of logit and probit conduct nonlinear maximum likelihood estimations in which the estimations of the coefficients for individual dummies and those for common independent variables are not independent. Therefore, the coefficients for common variables are also biased when the coefficients for individual dummy variables are estimated (Chamberlin, 1980; Hsiao, 1996).

<i>Variable</i>	<i>Explanation</i>
G_n	Normalized Greed: $\frac{T-R}{T-S}$
C_n	Normalized Cooperators' Gain: $\frac{R-P}{T-S}$
<i>Stage</i>	Index of stage in a phase, 1 to 12
<i>P2High</i>	A dummy variable that takes a value of 1 if an observation belongs to Phase 2 high cooperators' group, 0 otherwise
<i>P2Low</i>	A binary variable that takes a value of 1 if an observation belongs to Phase 2 low cooperators' group, 0 otherwise
<i>PrvDC</i>	A binary variable that takes a value of 1 if individual i defected and the partner cooperated in the previous stage, 0 otherwise
<i>PrvCD</i>	A binary variable that takes a value of 1 if individual i cooperated and the partner defected in the previous stage, 0 otherwise
<i>PrvCC</i>	A binary variable that takes a value of 1 if both individual i and his/her partner cooperated in the previous stage

Table 6.2: Variables and Explanation

fixed effects logit analyses that remove the heterogeneity term by conditioning the likelihood on the number of successes within a group – the number of cooperative choices made by each individual, when applied to the current data. The implementation of the conditional likelihood function, suggested by Chamberlain (1980), to our data yields

$$L^c = \prod_{i=1}^N \Pr(Y^{i1} = y^{i1}, Y^{i2} = y^{i2}, \dots, Y^{iT} = y^{iT} | \sum_{t=1}^T y^{it}), \quad (6.5)$$

in which N is the number of individuals, T is the number of observations that belong to the same individual, y^{it} is the dependent variable that takes a value of 1 when individual i cooperates at stage t , 0 otherwise.

Table 6.2 defines the variables used in the estimation of the models.⁵ Table 6.3 reports the results of the simple logit and conditional (fixed effects) logit estimations.⁶ In

⁵ F_n was not included since the sum of G_n , C_n , and F_n is always 1. When both *P2High* and *P2Low* are zeroes, the observation belongs to Phase 1. The baseline for the three binary variables related to the outcome of the previous stage is the case in which both players defected in the previous stage.

⁶The number of observations in the conditional logit model is smaller than that in the simple logit because

Variable	Logit		Conditional Logit	
	Coefficient	t	Coefficient	t
<i>ONE</i>	-3.745	3.64	-	-
<i>G_n</i>	1.362	0.82	1.161	0.68
<i>C_n</i>	5.773	3.36	5.245	2.97
<i>Stage</i>	-0.071	1.55	-0.088	1.84
<i>P2High</i>	-0.044	0.14	-1.069	2.34
<i>P2Low</i>	-0.932	2.64	0.477	0.97
<i>PrvDC</i>	1.285	3.36	0.832	1.97
<i>PrvCD</i>	1.214	3.16	0.426	1.02
<i>PrvCC</i>	4.713	11.66	3.581	8.07
# obs	528		484	
LR $\chi^2(8)$	279.04		138.86	
P > χ^2	0		0	
Pseudo R^2	0.4254		-	
ln L	-188.4		-137.8	

Table 6.3: Cooperation: Logit and Conditional Fixed Effects Logit Estimates

terms of the significance of the covariates, logit and fixed-effects logit models yield similar results.

First, the regression results indicate that the material payoff structure has a significant impact on behavior in the context of the repeated 2×2 social dilemma game. Ahn et al. (2001) report that the normalized values of fear, greed, and cooperators' gain have significant impacts on the players' strategy choice in the context of single play 2×2 social dilemma games. Whether this would still be the case when the material payoff structures vary across stages of a repeated game is an interesting empirical question. A simple tabulation of the frequencies of *Cooperation* across the six material payoff structures shown in Table 6.4 indicates that the choice of *Cooperation* is much more likely in the games with the high *Cooperators' Gain* (games 1, 3, and 5) than in those with the low *Cooperators'*

the observations from stages 1, and those belonging to the individuals whose choices do not vary across 22 stages, are all dropped.

	<i>High Cooperators' Gain</i>	<i>Low Cooperators' Gain</i>
<i>Greed = Fear</i>	Game 1 : 38 (39.6%) ($C_n = 0.4, G_n = 0.3, F_n = 0.3$)	Game 2 : 31 (32.3%) ($C_n = 0.2, G_n = 0.4, F_n = 0.4$)
<i>Greed < Fear</i>	Game 3 : 34 (35.4%) ($C_n = 0.4, G_n = 0.2, F_n = 0.4$)	Game 4 : 25 (26.0%) ($C_n = 0.2, G_n = 0.3, F_n = 0.5$)
<i>Greed > Fear</i>	Game 5 : 35 (36.5%) ($C_n = 0.4, G_n = 0.4, F_n = 0.2$)	Game 6 : 24 (25.0%) ($C_n = 0.2, G_n = 0.5, F_n = 0.3$)
<i>Total</i>	107 (31.2%)	80 (27.8%)

Table 6.4: The Frequency of Cooperation across Game Structures

Gain (games 2, 4, and 6). The table also suggests that when the value of *Cooperators' Gain* is fixed, the relative values of normalized *Fear* and *Greed* do not matter significantly.

The apparently significant difference between the frequencies of *Cooperation* in the games with the high and those with the low *Cooperators's Gain* is without controlling for the other factors in the repeated game. For example, the first stage in the repeated game has the high value of *Cooperators's Gain* while the last stage has the low value. It is possible that the apparent difference is due not to the value of the *Cooperators's Gain* but to the stage impact in the repeated game.

However, the large, as well as significant, coefficients for C_n in Table 6.3 indicate that the tabulation result is still valid when the impacts of other factors are controlled in regression analyses. It should be regarded as quite strong since the first stage with the high *Cooperators' Gain* is not included in the regressions due to the lack of variables related to the previous stage outcome.

Why does the material payoff structure still matter in the repeated game even after controlling for the impacts of the number of stages remaining and the outcome of the previous stage? One possibility is that there are players who treat each of the 12 stages not

as a stage in a repeated game but as an independent game. A more persuasive scenario is that the changes in the material payoff structures provide some players with the focal point for strategy switching. When a player in the [mutual *Defection*] equilibrium tries to send the signal to the partner of her intention to switch to the cooperative equilibrium, she can do that less costly when the *Cooperators's* gain is large and thus, the *Greed* and *Fear* are relatively small. For example, in Game 1314 shown in Figure 6.4, in which the two players switch to the cooperative equilibrium at exactly the same stage, the transition happens in a stage with the high *Cooperators' Gain*. In general, a closer look at Figures 6.4 to 6.12 reveals that the attempts to send cooperative signals usually occur in the stages with the higher *Cooperators' Gain*.

The collapse of cooperative plays at stages with the low *Cooperators' Gain* can be another contributing factor to the significantly different frequencies of *Cooperation* between the stages with the high and low *Cooperators' Gain*. However, a visual examination of the data does not clearly distinguish this hypothesis from the end-game effect, since the final stage has the low *Cooperators' Gain*.

Second, as the large and very significant coefficient for the variable *PrvCC* indicates, a player is much more likely to cooperate when she and the partner both cooperated in the previous stage. [Mutual *Cooperation*] is a sign that the two players are in the cooperative equilibrium and they have, regardless of their types, the incentive to sustain [mutual *Cooperation*] as long as possible. The variables', *PrvDC* and *PrvCD*, impacts are not robust. The coefficients for the two variables are significant in the simple logit model, but they are neither large nor very significant when the unobserved inter-individual heterogeneity is

controlled in the conditional fixed effect logit model.

The most interesting difference between the simple and the conditional logit estimations is found in the impacts of Phase 2 group variables, *P2High* and *P2Low*. To recall, in Phase 2 subjects belonged to either a *High* or *Low* group – simply called group 1 and group 2 during the experiment – depending on the number of cooperative choices they made in Phase 1. Therefore, in terms of group belongings, there are three possibilities: Phase 1, Phase 2 High group(*P2High*), and Phase 2 low group(*P2Low*). Two dummy variables *P2High* and *P2Low*, are included to assess the impacts of the Phase 2 group belongings. In the simple logit analysis, the variable *P2Low* has a significantly negative coefficient, while in the fixed-effects logit model, the variable *P2High* has a significantly negative coefficient.

At first glance, the result of the simple logit analysis is quite intuitive: individuals are less likely to cooperate when they belong to a low cooperators' group in Phase 2. However, the problem with the simple logit estimation is that it treats all the individuals homogeneously in terms of their baseline tendency to cooperate; the simple logit model regards the behavioral difference purely as the impacts of the variations in the dependent variables.

On the other hand, the result of the conditional logit estimation indicates that after players' types are controlled for, belonging to the Phase 2 high group has a negative impact on a player's likelihood of cooperating. To put it differently, a player in the Phase 2 high cooperators' group is less likely to match the number of cooperative choices he played in Phase 1. Table 6.5 presents this phenomenon at an individual level by tabulating the number of cooperative choices made in Phases 1 and 2 by the 12 individuals who belong

ID	Phase 1	Phase 2
13	8*	2
14	8*	9
16	2*	1
18	5	9*
23	9*	4
24	8	10*
27	5*	3
28	8	11*
31	12*	3
32	12*	2
37	3*	0
38	3*	2
Total	83	56

Table 6.5: Phase 2 High Group Individuals' Behavior in Phases 1 and 2

to the Phase 2 high cooperators' group. The table shows that only 3 out of 12 players who belonged to Phase 2 high cooperators' group made more cooperative choices in Phase 2 than they did in Phase 1 – for 9 players, the number of cooperative choices actually decreased from Phase 1 to Phase 2.

Table 6.6 provides another intuitive illustration of why the coefficients for Phase 2 group belonging variables differ between the simple and the conditional logit models and what aspects of the data each model represents. The table is constructed such that the 24 players are divided according to their Phase 2 belongings, and their frequencies of *Cooperation* are tabulated by phases. For example, cell *a* shows that the players who eventually belonged to Phase 2 high cooperators' group made 83(57.6%) cooperative choices in Phase 1.

The significantly negative coefficient for *P2Low* in the simple logit model represents the difference between the cells *c* and *d* – after taking into account the impacts of other variables but not the players' types. The nonsignificant coefficient for *P2High* in the

	<i>P2High</i>	<i>P2Low</i>	Total
<i>Phase 1</i>	83 (57.6%) ^a	14 (9.7%) ^e	97 (33.7%) ^c
<i>Phase 2</i>	56 (38.9%) ^b	34 (23.6%) ^d	90 (31.3%)
Total	139 (42.3%)	48 (16.7%)	187 (32.5%)

Table 6.6: Cooperation in Phases 1 and 2: By Phase 2 Group

simple logit model represents the small difference between the cells *b* and *c*. In the conditional fixed effects logit model, the significantly negative coefficient for *P2High* represents the large difference between the cells *a* and *b* – because the players’ types are accounted in the model, the reference point for cell *b* is no longer *c*, but *a*. Finally, the insignificantly positive coefficient for *P2Low* in the conditional fixed effects logit model represents the difference between cells *d* and *e*. The significance and magnitude of the coefficients for the four variables may differ slightly from what is intuitively shown in Table 6.5 because the coefficients are estimated while controlling for the impacts of the other variables, and because the observations from stage 1 are all dropped due to the absence of the values for the previous stage outcome variables.

Why do the players in the Phase 2 high cooperators’ group make less cooperative choices than they did in Phase 1? Coordination failure and type mismatch appear to provide a better account. Four players – players 13, 23, 31, and 32 – collectively made 30 less cooperative choices in Phase 2 than they did in Phase 1. Three of them – players 13, 31, and 32 – cooperate to the final stage of Phase 1 and start Phase 2 with *Cooperation*, indicating that they are most likely to be *Assurance* types. Player 23 starts both Phase 1 and Phase 2 with *Cooperation*, but he defects in the final stage of Phase 1, indicating that the player is either an *Assurance* type or a *PD* type who is strongly inclined to play the cooperative equilibrium. The combined cooperative choices made by these four players

dramatically decrease from 41 in Phase 1 and 11 in Phase 2, accounting for most of the difference between the number of cooperative choices made by players in the Phase 2 high cooperators' group in the two phases.

A closer look at their, and their partners', behavior in phases 1 and 2 indicates that the reason for the significant decline in the number of cooperative choices is most likely due to the coordination problem and type mismatch. For example, player 31 cooperates in all of the 12 stages of Phase 1 and in the first stage of Phase 2. But player 31's Phase 2 partner (player 37, who cooperated only twice in Phase 1) neither cooperates in stage 1, nor reciprocates player 32's *Cooperation* by cooperating in the next stage. Nevertheless, player 32 cooperates again in stage 3, inviting player 37 to the cooperative equilibrium. Again, player 37 does not reciprocate and player 32 eventually gives up any attempt to lead his partner to the cooperative equilibrium.

In retrospect, it was in part player 32's good luck that the player was able to enjoy mutual *Cooperation* for the entire 12 stages of Phase 1; his Phase 1 partner (player 31) was not only an *Assurance* type, but also willing to try the cooperative equilibrium strategy. Not every *Assurance* type is so lucky. For example, player 22, whose Phase 2 behavior strongly suggests that he is of the *Assurance* type, met a player who refused to reciprocate player 22's *Cooperation* in both stages 1 and 3. Player 22 had no choice but to convert to *Defection* himself, scoring *Cooperation* only twice and ending up in the low cooperators' group in Phase 2.

In general, when an individual has chosen a relatively large number of *Cooperation* in Phase 1, part of the reason is that the player was lucky; being cooperative and ready to

take the risk of playing cooperative equilibrium is a necessary but not sufficient condition for a player to actually play many cooperative choices. He also needs to be lucky enough to meet a partner who is either cooperative or at least wise enough to behave reciprocally.

As the principle of “regression towards the mean” suggests, an individual who is quite lucky today is less likely to be that lucky again tomorrow. So, a player with a large number of Phase 1 *Cooperation* may find it difficult to play the same, or a higher, level of cooperative choices in Phase 2. Especially, when there are some individuals in Phase 2 high cooperators’ group whose *Cooperation* in Phase 1 reflects not their type but their strategic consideration for Phase 2 group belonging, the *Assurance*-type players with higher numbers of Phase 1 *Cooperation* will have difficulty in maintaining the same level of cooperative choices in Phase 2.⁷

6.5 Interdependence of Strategies: Bivariate Probit Analyses

This section reports a series of bivariate probit analyses of the experimental data. In the finitely repeated 2×2 social dilemma game, the choices of any two paired players at any stage are expected to be highly correlated. Since any two paired players’ choices are not independent, simple binary choice models do not exactly reflect the underlying data generation process. The bivariate probit model utilizes the bivariate normal distribution function to model the correlation between two interdependent binary choice variables. However,

⁷In addition, the two most cooperative players in Phase 1 of any session were not matched with each other in Phase 2, increasing the chance that their Phase 2 partners are not very cooperative types.

		Column Player		Total
		<i>Defection</i>	<i>Cooperation</i>	
Row Player	<i>Defection</i>	55.9 %	11.8 %	195
	<i>Cooperation</i>	11.5 %	20.8 %	93
	Total	194	94	288

Table 6.7: Correlation between Paired Players' Choices

while the logistic function eliminates individual heterogeneity terms in the conditional likelihood function, the normal distribution function does not. Therefore, it is not possible to take account both the problems of heterogeneity and interdependence in a single regression model.⁸

Therefore, the series of bivariate probit analyses in this section have a narrower focus; in addition to examining whether the basic results of the fixed effects logit estimations still hold when the problem of interdependence is taken into account, the empirical analyses of this section focus on identifying the interdependence and its source.

Table 6.7 tabulates the choices made by all the paired subjects. The simple correlation coefficient between the choices of the row players and column players is quite high at 0.47. In 73% of a total of 288 stages, two players choose the same strategy. Since there is a considerable proportion of cases in which both players choose *Cooperation*, the high correlation cannot be attributed simply to the prevalence of noncooperative choices in the data.

The equilibrium analyses of Chapter 5 have provided an explanation why the choices of two paired players at a given stage are highly correlated. Any type of equilibrium

⁸It is possible to include individual dummy variables in a bivariate probit model so that both the problems of interdependence and heterogeneity can be accounted for. However, including individual dummy variables can cause bias in estimates when the number of observations for each individual is not very large. This section tries the model later in an exploratory manner.

involves the same choice by both of the players, except for the stage in the cooperative equilibrium in which behavior of different types diverges. When the two paired players are already in equilibrium, they know which equilibrium they are in by the way the game is played up to the current stage. On the other hand, when the players have been engaged in non-equilibrium signal exchanges in the previous stages, their choices in the current stage reflect the (im)patience of the signal sender and the (non)responsiveness of the signal receiver. In any case, the way the game is played in the immediately preceding stage is the simplest focal point on which the two paired players can coordinate their choices in the current stage.

Drawing on the random utility model (6.2), let $u_r^{i,t}$ ($u_c^{i,t}$) denote the difference between the expected utilities of *Cooperation* and *Defection* for the row(column) player at stage t of the repeated game played by i th pair:

$$u_r^{i,t} = X_r^i \beta_r + \epsilon_r, \quad i = 1, 2, \dots, N$$

$$u_c^{i,t} = X_c^i \beta_c + \epsilon_c, \quad i = 1, 2, \dots, N.$$

Our observation is

$$\begin{aligned} (y_r^{i,t}, y_c^{i,t}) &= (1, 1) && \text{if } u_r^{i,t} > 0 \text{ and } u_c^{i,t} > 0 \\ &= (1, 0) && \text{if } u_r^{i,t} > 0 \text{ and } u_c^{i,t} \leq 0 \\ &= (0, 1) && \text{if } u_r^{i,t} \leq 0 \text{ and } u_c^{i,t} > 0 \end{aligned}$$

$$= (0, 0) \quad \text{if } u_r^{i,t} \leq 0 \text{ and } u_c^{i,t} \leq 0.$$

The bivariate probit model estimates parameters β_r , β_c , and ρ , in the likelihood function:

$$L = \prod_{i=1}^N \int_{\tau_r}^{\lambda_r} \int_{\tau_c}^{\lambda_c} \phi_2(\epsilon_r, \epsilon_c; \rho) d\epsilon_r d\epsilon_c, \quad (6.6)$$

where ϕ_2 , the bivariate normal density function, is

$$\phi_2(\epsilon_r, \epsilon_c; \rho) = [2\pi(1 - \rho^2)^{\frac{1}{2}}]^{-1} \exp[-\frac{1}{2}(1 - \rho^2)^{-1}(\epsilon_r^2 + \epsilon_c^2 - 2\rho\epsilon_r\epsilon_c)] \quad (6.7)$$

with $E(\epsilon_r) = E(\epsilon_c) = 0$, $Var(\epsilon_r^2) = Var(\epsilon_c^2) = 1$, and ρ as the correlation coefficient. The limits of integration depend on the observed choices of the two players $(y_r^{i,t}, y_c^{i,t})$ as follows:

$$\text{If } (y_r^{i,t}, y_c^{i,t}) = (0, 0), \lambda_r = -X_r\beta_r, \tau_r = -\infty, \lambda_c = -X_c\beta_c, \tau_c = -\infty.$$

$$\text{If } (y_r^{i,t}, y_c^{i,t}) = (1, 0), \lambda_r = \infty, \tau_r = -X_r\beta_r, \lambda_c = -X_c\beta_c, \tau_c = -\infty.$$

$$\text{If } (y_r^{i,t}, y_c^{i,t}) = (0, 1), \lambda_r = -X_r\beta_r, \tau_r = -\infty, \lambda_c = \infty, \tau_c = -X_c\beta_c.$$

$$\text{If } (y_r^{i,t}, y_c^{i,t}) = (1, 1), \lambda_r = \infty, \tau_r = -X_r\beta_r, \lambda_c = \infty, \tau_c = -X_c\beta_c.$$

X_r and X_c are row vectors of exogenous variables that determine the expected utilities of *Cooperation* and *Defection*.

Because the division of the subjects into the row and column players is an arbitrary labeling, this study uses a model of pooled bivariate probit that restricts the coefficients

Variable	Model 1		Model 2		Model 3	
	Coefficient	t	Coefficient	t	Coefficient	t
<i>ONE</i>	-1.947	3.69	-0.837	2.15	-0.875	4.48
<i>G_n</i>	0.376	0.43	0.215	0.31		
<i>C_n</i>	2.964	3.30	1.266	1.87		
<i>Stage</i>	-0.046	1.92	-0.006	0.33	-0.430	1.91
<i>P2High</i>	-0.076	0.45	0.090	0.66	-0.580	0.35
<i>P2Low</i>	-0.330	1.74	-0.339	2.29	-0.353	1.90
<i>PrvDC</i>	0.693	3.25			0.536	2.64
<i>PrvCD</i>	0.751	3.56			0.585	2.90
<i>PrvCC</i>	2.721	13.42			2.584	13.47
<hr/>						
# Obs	528		528		528	
- ln L	385		578		398	
Wald χ^2	338.56		15.49		329.95	
P > χ^2	0.000		0.115		0.000	
ρ	0.229		0.758		0.290	
LR(ρ) χ^2	4.87		147.21		8.63	
P > χ^2	0.027		0.000		0.003	

Table 6.8: Cooperation: Pooled Bivariate Probit Estimates

for the exogenous variables the same between the equations for the row and column players while efficiently utilizing the data.⁹

Table 6.8 reports the results from a series of pooled bivariate probit estimations.

Model 1 uses a specification identical to that in the previous logit estimations. In terms

⁹An observation in the bivariate probit model has a structure

$$y_{1i}, X_{1i}, y_{2i}, X_{2i}$$

in which y_{1i} and y_{2i} are the two dependent binary variables and X_1 (X_{2i}) is the vector of covariates of y_{1i} (y_{2i}).

In the experimental data, the division of the players into row and column players is arbitrary. While the experimental computer program divided the subjects into row and column players for the purpose of recording, all the subjects saw themselves as the row players in their own computer monitor. There is no reason, except for the stochastic effect, to hypothesize that the impacts of covariates differ between the choices of the arbitrarily labeled row and column players. A convenient way to achieve the same coefficients for row and column players is to enter each observation twice as follows:

$$y_{1i}, X_{1i}, y_{2i}, X_{2i}$$

$$y_{2i}, X_{2i}, y_{1i}, X_{1i}$$

of the magnitude and significance of the coefficient estimates, the result is substantively identical to that of the logit estimation. Model 2 and Model 3 are included to demonstrate the source of the correlation between the choices of *Row* and *Column* players. In the bivariate probit model, the estimate of correlation coefficient ρ captures the correlation between two binary dependent variables' responses to the random shock – the influence of all the factors not included as independent variables. Therefore, in principle, when all the factors that have impacts on the two dependent variables are perfectly specified in the model, the correlation coefficient should be 0.

By alternately excluding each set of variables from the model specification and comparing the resulting estimates of the correlation coefficient across different specifications, we can see which variables are the biggest source of the apparent correlation between the two paired players' choices.

When the variables representing the outcome of the previous stage are excluded (Model 2), the correlation coefficient ρ becomes very large (0.758) and extremely significant. On the other hand, when the variables related to the normalized values of the payoff parameters are excluded from the estimation (Model 3), the estimate of correlation coefficient does not change significantly from that in Model 1. The estimate of the correlation coefficient does not differ significantly from that in Model 1, when other sets of independent variables are excluded from the estimation: 0.233 when *P2High* and *P2Low* are excluded, and 0.245 when *stage* is excluded. Compared to the estimated correlation coefficient in Model 1, this implies that the outcome of the previous stage is the most important source

of the apparent correlation between the two paired players' strategies.¹⁰

Finally, Table 6.9 reports a pooled bivariate probit estimation that includes individual dummies as independent variables. As discussed earlier, variants of the probit model do not allow the elimination of individual heterogeneity in the conditional likelihood function, and the inclusion of individual dummies may bias the whole estimation when T , the number of observations for each individual, is small. Therefore, the estimation shown in Table 6.9 is only an exploratory attempt to compare its results with those of the fixed effects logit and bivariate probit models.

Table 6.9 does show some interesting results. First, the variables C_n and $PrvCC$ have the largest and the most significant coefficients. Second, as was the case in the fixed effects logit model, the coefficient for $P2High$ has a significantly negative sign. Third, after the heterogeneity is controlled, the correlation coefficient ρ becomes indistinguishable from 0. In general, the pooled bivariate probit model including individual dummy variables, in spite of its theoretical weakness, confirms the major empirical results of the fixed effects logit and bivariate probit estimations.

6.6 Conclusion

This chapter tested the equilibrium analyses of Chapter 5 using an experimental data set of the finitely repeated 2×2 social dilemma game. The strict equilibrium plays were rare, but each of the three broad types of the sequential equilibrium – [all *Defection*] equilibrium, cooperative equilibrium, and hybrid equilibrium – appears at least once in

¹⁰Note that even in Model 1, ρ is estimated to be significantly larger than 0. This implies that two paired players tend to respond to the factors not included in the regression in the same direction.

Variable	Coefficients	t
<i>ONE</i>	-1.359	2.00
<i>G_n</i>	0.617	0.66
<i>C_n</i>	3.162	3.31
<i>stage</i>	-0.056	2.17
<i>P2High</i>	-0.445	2.22
<i>P2Low</i>	-0.065	0.27
<i>PrvDC</i>	0.502	2.10
<i>PrvCD</i>	0.360	1.52
<i>PrvCC</i>	2.232	9.36
<hr/>		
# Obs	528	
-ln L	348	
Wald $\chi^2(62)$	347.64	
P> χ^2	0.000	
ρ	0.103	
LR(ρ) χ^2	0.81	
P> χ^2	0.368	

*Estimates of the coefficients for individual dummies are not reported

Table 6.9: Cooperation: Pooled Bivariate Probit Estimates with Individual Dummies

the data. Some subjects do cooperate to the final stage of the repeated game, and others patiently try to invite the partner to the cooperative equilibrium.

A series of the fixed effect logit and the bivariate probit models are estimated to test the impacts of the theoretical variables systematically. The results of the statistical models indicate that an individual has the highest probability of cooperating when both the individual and the partner cooperated in the immediately preceding stage. The probability of *Cooperation* is shown to be significantly higher, across all the statistical models, in the stages in which the material gains from mutual *Cooperation* are high.

A noticeable result of the fixed effects logit analysis is that individuals who made more cooperative choices in Phase 1 are not likely to match the level in Phase 2, even when they belong to the high cooperators' group. This result is interpreted as indicating the

importance of the coordination and partner's type in facilitating an individual's cooperative behavior. The bivariate probit models confirm the general importance of the outcome of the previous stage and the material payoff structure of the current stage in determining a player's choice. In addition, by comparing the estimates of the correlation coefficient across different specifications of the bivariate probit model, we were able to see that the biggest source of the interdependence between the two paired players' choices is the way the game is played in the previous stage.

Chapter 7

Conclusion: Heterogeneous Motivations and Cooperation in Social Dilemmas

This dissertation has attempted to incorporate heterogeneous motivations into game theoretic models of social dilemmas. When a group of rational individuals holding heterogeneous motivations interact in a social dilemma, what outcomes would result? How do the material, institutional, and cultural factors affect individuals' behavior and social outcomes? This chapter summarizes theoretical and empirical findings of this study and provides a few suggestions for further research.

In Chapter 1, I reviewed several extant institutional approaches to social dilemmas. The first set of institutional approaches includes the State solution and the Market/Privatization solution. Examples of these are the intervention by external authorities

with punishment and reward rules, selective incentives provided for the cooperators, and dissolution of the dilemma structure via market/privatization. In any case, individuals face a new incentive structure that is not a dilemma. However, each of these solutions has its own problems.

There is no guarantee that the external authority would implement the rules in a fair manner. More often than not, even when an external authority does intend to fairly implement the rules, it may not have the proper resources and information necessary for a successful intervention. The Market/Privatization solution can be very costly, and it often forgoes the social opportunity of utilizing the economy of scale embedded in social dilemmas.

Self-governance is the second institutional approach to explaining and facilitating cooperation in social dilemmas. Rather than being inexorably trapped in the dilemma, individuals can utilize the local knowledge they have, devise their own rules, make commitments to each other, and preserve the productive social opportunity without relying on external authorities. Theoretical and empirical work on self-governance have requested a fundamental shift in the ways policymakers think of the solutions to social dilemmas.

This dissertation, while acknowledging the importance of institutional factors, tried to incorporate another fundamental factor – individual motivations – to the study of social dilemmas. An institution is a function that transforms initial material and cultural conditions to aggregate social outcomes. Institutions are designed to solve social problems by channelling individuals' behavior to desired social outcomes. Motivations – defined in this study as individuals' preference over the alternative social outcomes that can result

from the initial conditions of an action situation – in turn, decide how individuals would react to the incentives provided by a specific institutional arrangement.

Institutions based on an invalid assumption regarding individual motivations would not be able to achieve the social goals for which they are designed. In addition, designing and maintaining institutions involve costs. How to direct the limited institutional resources efficiently also depends on an understanding of motivations that individuals hold. Moreover, institutions not only alter immediate behavior but reshape, in the long run, the norms and culture of a society.

Simply assuming that all individuals are selfish is not only empirically invalid, but also can result in seriously detrimental policy prescriptions. The universal selfishness assumption, when used as the basis for policy prescription, can undermine the productive social opportunity embedded in social dilemmas. It can also be detrimental to the development of social norms to cope with the problem of social dilemmas.

On the empirical side, enough evidence has been accumulated showing that, in social dilemmas in particular, there is a significant proportion of individuals for whom the purpose of action is not to achieve the maximum material gains for themselves. However, efforts to develop formal models of social dilemmas that take into account the heterogeneous motivations have only recently begun.

The key question in providing an alternative formal model that replaces those based on the universal selfishness assumption is how to model heterogeneity. That is, how to incorporate in a formal model the fact that some individuals are selfish, some are not, and the non-selfish individuals also differ among themselves in terms of the extent to which

they depart from a purely selfish motivation.

Well-known, quasi-game theoretic models of heterogeneous motivations do exist. Evolutionary models of social dilemmas have examined the conditions under which agents with reciprocity principles can prosper (Axelrod, 1981; Axelrod and Hamilton, 1981; Bendor and Swistak, 1997). However, the standard evolutionary models do not exactly capture the rationality of human decision making. In evolutionary models, a type is defined by the strategy it uses. But a rational human actor with non-selfish motivation would still exhibit varying behavior responding to material, institutional, and population conditions. While the evolutionary models have contributed greatly in exploring the conditions for the survival and proliferation of reciprocal behavior, a further development is necessary in which individuals are modeled to have rational preference and make flexible decisions.

The standard game theoretic models of cooperation in the finitely repeated Prisoner's Dilemma (Kreps et al., 1982; Fudenberg and Maskin, 1986) appear to provide rationale for cooperative behavior in social dilemmas. However, they provide rationality of cooperation only for the selfish individuals, while labeling the non-selfish individuals "irrational" or even "crazy." Non-selfish motivations are marginalized in those standard game theoretic models.

This dissertation, drawing on the recent development of behavioral game theory, incorporated both the rationality of individuals and heterogeneity across individuals into a coherent game theoretic model of a social dilemma. Developing the model required, first of all, a behavioral(empirical) definition of a social dilemma that is based on the observable factors of the action situation. Second, a behavioral framework of social dilemma has

been developed after critically reviewing two exemplary frameworks: Vernon Smith's MES (microeconomic system) framework and the IAD (institutional analysis and development) framework. The behavioral framework connects the empirical world of social dilemmas and the deductive formal models. It also makes it explicit how, and in what sense, a formal model provides an explanation for an empirical question and what assumptions are necessary for the explanation to be valid.

Chapter 4 was an exercise in comparing theoretical and empirical performance of alternative motivational models. The models of altruism and inequity aversion, which have been considered as promising alternatives to the model of universal selfishness, were investigated. When applied to 2×2 social dilemma games, each of the two models generates a series of preference-ordering types over the four outcomes of the game, while the traditional assumption of self-interest allows only one preference type. The model of altruism classifies individuals into four preference-ordering types, but the possibilities of certain types are limited by the structure of material payoffs of a game. On the other hand, the model of inequity aversion generates only two preference-ordering types and their existence is not limited by the structure of the material payoffs. The altruism model predicts that there are individuals who cooperate no matter what the other individual does. The model of inequity aversion, on the other hand, precludes unconditional cooperation and divides individuals into two broad subsets of conditional cooperators and unconditional defectors.

When the social dilemma is framed as an incomplete information game, both the models specify conditions for cooperative equilibrium in simultaneous and sequential 2×2 social dilemma games. The equilibrium analysis allows us to derive hypotheses regarding

the relative frequency of cooperation in the four qualitatively different information sets of the games. The hypotheses based on the altruism model are less restrictive than those that are based on the model of inequity aversion.

Empirical tests are conducted drawing on two sets of experimental data. In terms of preference ordering, the model of inequity aversion accounts for a substantive proportion of the preference types not explained by the pure selfishness model. In contrast, the altruism model does not provide meaningful additional explanation for the types that are not accounted by the inequity aversion model. In terms of the behavior in the four qualitatively different information sets, the data strongly supports the hypotheses based on the model of inequity aversion.

Chapter 5 has studied equilibria of the finitely repeated 2×2 social dilemma game while modeling heterogeneity among individuals based on the empirical conclusions of Chapter 4. By modeling the non-selfish individuals as well as the selfish ones as rational decision makers, the analysis provided a direct comparison to the standard finitely repeated *Prisoner's Dilemma* model. The equilibrium analyses are conducted first, assuming that there are only two types and second, assuming that there is a continuum of types. In both cases, there does exist the cooperative equilibrium of the finitely repeated game whenever there exists the stage game cooperative equilibrium.

When there does not exist the stage game cooperative equilibrium, the possibility of cooperation in the finitely repeated game depends on the *Defection* point of the *PD*-type players when the *Assurance* types use the grim trigger strategy, and there being multiple *Defection* points among the *PD* types that facilitate the *Assurance* types' use of the grim

trigger strategy. There also exist a series of hybrid equilibria in which a transition from mutual *Defection* to mutual *Cooperation* occurs. Though the exact replication of this kind of equilibria in real settings is not very likely, it still provides a rational basis for the risky investment/initiation by the players who want to escape from the trap of mutual *Defection*. As was the case in the analyses of the static models in Chapter 4, the possibility of cooperation is affected by the environment of an action situation – the material payoff structure of a 2×2 social dilemma – as well as the culture of a group – distribution of types within the population. In addition, with the existence of multiple equilibria, the problem of coordination becomes more significant.

Chapter 6 conducted a series of empirical tests of the equilibrium analyses of Chapter 5 using an experimental data set. The strict equilibrium plays were rare, but each of the three broad types of the sequential equilibrium – [all *Defection*] equilibrium, cooperative equilibrium, and hybrid equilibrium – appears at least once in the data. Some subjects do cooperate to the final stage of the repeated game, and others patiently try to invite the partner to the cooperative equilibrium.

A series of the fixed effects logit and the bivariate probit models are estimated to test the impacts of the theoretical variables systematically. The results of the statistical models indicate that an individual has the highest probability of cooperating when both the individual and the partner cooperated in the immediately preceding stage. The probability of *Cooperation* is shown to be significantly higher, across all the statistical models, in the stages in which the material gains from mutual *Cooperation* are high.

A noticeable result of the fixed effects logit analysis is that individuals who made

more cooperative choices in Phase 1 are not likely to match the level in Phase 2, even when they belong to the high cooperators' group. This result is interpreted as indicating the importance of the coordination and partner's type in facilitating an individual's cooperative behavior. The bivariate probit models confirm the general importance of the outcome of the previous stage and the material payoff structure of the current stage in determining a player's choice. In addition, by comparing the estimates of the correlation coefficient across different specifications of the bivariate probit model, we were able to see that the biggest source of the interdependence between the two paired players' choices is the way the game is played in the previous stage.

Suggestions for Further Research

While this dissertation emphasized the importance of incorporating multiple motivations in game theoretic models of social dilemmas, the actual theoretical analyses focused more on developing benchmark models of social dilemmas under minimal institutional settings: the basic simultaneous 2×2 social dilemmas. The analyses of sequential and finitely repeated games can be viewed as extensions of the basic model to institutionally richer settings; sequential interactions and repeated interactions are often results of a conscious choice by individuals involved in social interactions. Recall that, unless heterogeneous motivations are incorporated, these modifications of the basic simultaneous social dilemma do not make any difference. But as the analyses in Chapters 4 and 5 have shown, even these very modest modifications do make meaningful differences: both the sequential and finitely repeated play of a social dilemma enhance the possibility of cooperation.

Given the analytical difference that the incorporation of heterogeneous motivations makes, it is necessary to extend the mode of analysis to incorporate diverse rules and institutions. For example, it will be interesting to see how rules that allow positive and negative subsidies, often called reward and sanction, would affect individuals' behavior and the social outcomes in the presence of multiple types of individuals. Fehr and Schmidt (1999) show that non-selfish types – those who prefer equitable distribution – may punish defectors, even when the punishment involves costs to the punishers. Knowing this, the selfish types may also cooperate. We have, then, a possibility for achieving mutual cooperation by appending a punishment phase to a single stage, simultaneous social dilemma game that does not have a cooperative equilibrium. We could also devise a finitely repeated social dilemma without a cooperative equilibrium and incorporate punishment rules to achieve a cooperative outcome.

Punishment or sanction, of course, is only one of the diverse institutional arrangements that are available to individuals to cope with the problem of social dilemmas. Would reward, in the form of a horizontal, positive subsidy, have the same beneficial effect? What if sanctioning and rewarding are implemented by government? Would it still have the same effect as when they are implemented horizontally, by individuals themselves involved in a social dilemma? How would the ways costs of sanctioning and reward are borne affect individuals' behavior? Given the wide utilization of reward and sanction in public policy as well as in self-governing institutions, more rigorous game theoretic models need to be developed to provide institutional analyses of these methods.

Rules related to information transmission provide another important area for

which game theoretic models with heterogeneous motivations need to be developed. The cooperation-enhancing effect of sequential dilemma games, in comparison with simultaneous games, critically depends on the fact that the motivational type of a first mover is revealed to a second mover. In general, we can safely hypothesize that rules that facilitate transmission of information regarding players' types to their current or potential partners in social interactions would enhance the possibility of cooperation. But social dilemmas take diverse forms and so are the ways in which information regarding a player's type can be transmitted to others. Therefore, it is necessary to address this question in a more rigorous way to generalize the mode of information transmission that enhances cooperation and the conditions under which information transmission becomes more effective.

Finally, when we acknowledge the existence of heterogeneous motivations across individuals, we have to address a fundamental question: Where does motivational heterogeneity come from? Why are some individuals selfish and others are not? Why do some groups have more non-selfish individuals than other groups? Evolutionary game theoretic models (Maynard Smith, 1982; Axelrod, 1981) seem to provide ways to address these questions. But those models are limited by their biological origin in their applicability to social dilemmas with rational human agents. The key problem of the standard evolutionary game theoretic models is that the heterogeneity across individuals is modeled in deterministic behavioral strategies. On the other hand, the models developed in this dissertation have shown that individuals, regardless of their motivations, can adjust their behavior to the subtleties of an action situation. For example, defecting in every stage is a way to model very selfish players in the standard evolutionary game theory. But in the models devel-

oped in this dissertation, even the purely selfish players do not always defect; they may cooperate in a sequential social dilemma when there is a high enough probability that the second mover is a conditional cooperator. They may also cooperate in a finitely repeated social dilemma game. Non-cooperative game theoretic models have often been criticized for assuming a hyper-rationality. However, compared to the deterministic strategy in the standard evolutionary game theoretic models, the rationality of human actors captured in the non-cooperative game theoretic models better reflect reality.

The strategic inflexibility of standard evolutionary models is especially problematic in analyzing the evolutionary impact of alternative institutions. Suppose there is a type of social dilemma with a minimal institutional setting, and a theorist wants to assess the effects of two alternative institutional arrangements in coping with the social dilemma. The theorist also wants to see the evolutionary consequences of the alternative institutions. The problem is how to model the strategies of a type across these institutional settings. Were we interested in animal behavior, it would be plausible to assume that the behavior of a type remains the same.¹ A more realistic assumption in analyzing human behavior is that each type of individual will respond to the institutional environment by adjusting their strategies.

The indirect evolutionary approach, a recent trend in evolutionary game theory, seems to be a promising way to model evolutionary processes with rational actors (Güth and Yaari, 1992; Güth, Kliemt, and Peleg, 2000; Bohnet, Frey, and Huck, 2001). In an indirect evolutionary model, the type of an individual is modeled in a utility function, in a

¹In fact, animals also exhibit rather sophisticated behavioral adjustment to environmental and population conditions. Therefore, strictly speaking, assuming an inflexible behavior is not realistic even for animals.

similar way to that used in the models developed in this dissertation. In an evolutionary process, the increase or decrease of a type depends not on the utility payoff but on the material payoff the type obtains. At any given stage of an evolutionary process, therefore, the behavior of players of different types is analyzed using the solution concepts of standard game theory that reflect human decision-making processes better than the deterministic strategies of standard evolutionary game theory. The indirect evolutionary approach provides a promising way to combine standard non-cooperative game theory and evolutionary game theory.

Bibliography

- Ahn, T. K., Elinor. Ostrom, David. Schmidt, Robert Shupp, and James. Walker. 2001. "Cooperation in PD Games: Fear, Greed, and History of Play." *Public Choice* 106(1/2):137-155.
- Ahn, T. K., Elinor. Ostrom, and James. Walker. 2000. "Altruism or Equity?: Modeling Non-Selfishness in Social Dilemmas." Bloomington: Indiana University, Workshop in Political Theory and Policy Analysis, Working Paper #W98-34.
- Alchian, Armen A. 1950. "Uncertainty, Evolution, and Economic Theory." *Journal of Political Economy* 58:211-221.
- Alchian, Arman A. and Harold Demsetz.1972. "Production, Information Costs, and Economic Organization." *American Economic Review* 77:388-401.
- Andreoni, James. 1990. "Impure Altruism and Donation to Public Goods: A Theory of Warm Glow Giving." *Economic Journal* 100:464-477.
- Andreoni, James, and Rachel Croson. Forthcoming. "Partners versus Strangers: Random Rematching in Public Goods Experiments." In *Handbook of Experimental Economics Results*, ed. Vernon L. Smith and Charles R. Plott. North Holland: Amsterdam.
- Andreoni, James, and John H. Miller. 1993. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence." *Economic Journal* 103: 570-585.

- Andreoni, James, and Hal Varian. 1999. "Preplay Contracting in the Prisoners' Dilemma." *Proc. Natl Acad.Sci.* 96:10933-10938.
- Aristotle. 1962. *Nichmachean Ethics*. Indianapolis: Bobbs-Merrill.
- Aumann, Robert. 1981. "Survey of Repeated Games." In *Essays in Game Theory and Mathematical Economics in Honor of Oscar Morgenstern*, ed. Robert Aumann. Mannheim: Bibliographisches Institut.
- Axelrod, R. 1981. "The Emergence of Cooperation among Egoists." *American Political Science Review* 75:306-318.
- Axelrod, R. and W. D. Hamilton. 1981. "The Evolution of Cooperation." *Science* 211:1390-1396.
- Bendor, Jonathan and Piotr Swistak. 1997. "The Evolutionary Stability of Cooperation." *American Political Science Review* 91:290-307.
- Benoit, Jean-Pierre, and Vijay Krishna. 1985. "Finitely Repeated Games." *Econometrica* 53:905-922.
- Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10:122-142.
- Binmore, Ken G. 1998. "The Evolution of Fairness Norms." *Rationality and Society* 10: 275-301.
- Bohnet, Iris, and Bruno S. Frey. 1998. "The Sound of Silence in Prisoner's Dilemma and Dictator Games." Berkeley: University of California, Haas School of Business,

Working Paper.

Bohnet, Iris, Bruno S. Frey, and Stephen Huck. 2001. "More Order with Less Law: On Contract Enforcement, Trust, and Crowding." *American Political Science Review* 95: 131-144.

Bolton, Gary E. 1991. "A Comparative Model of Bargaining: Theory and Evidence." *American Economic Review* 81:1096-1136.

Bolton, Gary E., Jordi Brandts, Elena Kaatok, Axel Ockenfels, and Rami Zwick. Forthcoming. "Testing Theories of Other-Regarding Behavior: A Sequence of Four laboratory Studies." In *Handbook of Experimental Economics Results*, ed. Vernon L. Smith and Charles R. Plott. North Holland: Amsterdam.

Bolton, Gary E., and Elena Katok. 1998. "An Experimental Test of the Crowding Out Hypothesis: The Nature of Beneficent Behavior." *Journal of Economic Behavior and Organization* 37:315-331.

Bolton, Gary E., Elena Katok, and Rami Zwick. 1998. "Dictator Game Giving: Rules of Fairness versus Acts of Kindness." *International Journal of Game Theory* 27:269-299.

Bolton, Gary E. and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90:166-193.

Bowles, Samuel. 1985. "The Production Process in a Competitive Economy: Walasian, neo-Hobbesian, and Marxian Models." *American Economic Review* 75:16-36.

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets

- and Other Economic Institutions." *Journal of Economic Literature* 36:75-111.
- Bowles, Samuel. 1999. "Individual Interactions, Group Conflicts and the Evolution of Preferences." In *Social Dynamics*. Steven Durlauf and Peyton Young, Cambridge: MIT Press, forthcoming.
- Brennan, Geoffrey, Werner Güth, and Hartmut Kliemt. 1997. "Trust in the Shadow of the Court if Judges Are No Better." Working Paper. Tilburg: Center for Economic Research, Tilburg University.
- Cain, Michael. 1998. "An Experimental Investigation of Motives and Information in the Prisoner's Dilemma Game." *Advances in Group Processes* 15: 133-160.
- Camerer, Colin. 1990. "Behavior Game Theory." In *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, ed. Robin Hogarth, 311-336. Chicago: University of Chicago Press.
- Camerer, Colin. 1995. "Individual Decision Making." In *The Handbook of Experimental Economics*, ed. John H Kagel and Alvin E. Roth, 587-703. Princeton, NJ: Princeton University Press.
- Camerer, Colin. 1997. "Progress in Behavioral Game Theory." *Journal of Economic Perspectives* 11(4):167-188.
- Camerer, Colin, and Richard Thaler. 1995. "Anomalies: Ultimatums, Dictators, and Manners." *Journal of Economic Perspectives* 9(2):209-219.
- Cameron, A. Colin and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data*.

Cambridge, MA: Cambridge University Press.

- Cameron, Lisa. 1995. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." Discussion Paper. Princeton, NJ: Princeton University.
- Carpenter, Jeffrey P. 1999. "Bargaining Outcomes as the Result of Coordinated Expectations: An Experimental Study of Sequential Bargaining." Working Paper. Middlebury, VT: Middlebury College.
- Carpenter, Jeffrey P. 1999. "Is Fairness Used Instrumentally?" Working Paper. Middlebury, VT: Middlebury College.
- Carpenter, Jeffrey P. Forthcoming. "Evolutionary Models of Bargaining: A Comparison by Means of Simulation." *Computational Economics*.
- Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47:225-238.
- Cho, In-Koo, and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics* 102:179-222.
- Cho, Kisuk, and Byoung-il Choi. 1999. "A Cross-Society Study of Trust and Reciprocity: Korea, Japan, and the U.S." Working Paper. Seoul: Ewha Women's University.
- Chong, Dennis. 1991 *Collective Action and the Civil Rights Movement*. Chicago: University of Chicago Press.
- Clark, Kenneth, and Martin Sefton. 2001. "The Sequential Prisoner's Dilemma: Evidence on Reciprocation." *Economic Journal* 111:51-68.

- Coleman, James. 1988. "Free Riders and Zealots: The Role of Social Networks." *Sociological Theory* 6:52-57.
- Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. 1996. "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games." *Games and Economic Behavior* 12:187-218.
- Crawford, Sue and Elinor Ostrom. 1995. "A Grammar of Institutions." *American Political Science Review* 89(3):582-600.
- Croson, Rachel T.A. 1998. "Theories of Altruism and Reciprocity: Evidence from Linear Public Goods Games." Working Paper. Philadelphia: Wharton School of Management, University of Pennsylvania.
- Davis, Douglas D., and Charles A. Holt. 1993. *Experimental Economics*. Princeton, NJ: Princeton University Press.
- Dawes, Robyn M. 1975. "Formal Models of Dilemmas in Social Decision Making." In *Human Judgment and Decision Processes: Mathematical Approaches*, ed. Martin F. Kaplan and Steven Schwartz, 87-108. New York: Academic Press.
- Dawes, Robyn M. 1980. "Social Dilemmas." *Annual Review of Psychology* 31:161-193.
- Dawes, Robyn M., Jeanne McTavish, and Harriet Shaklee. 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.
- Demsetz, Harold. 1997. "The Primacy of Economics: An Explanation of the Compar-

- ative Success of Economics in the Social Sciences.” Western Economic Association International 1996 Presidential Address. *Economic Inquiry* 35:1-11.
- Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.
- Dougherty, Keith L. and Michael J.G. Cain. 1999. “Linear Altruism and the 2×2 Prisoner’s Dilemma.” Working Paper. Miami: Florida International University.
- Dufwenberg, Martin, and Georg Kirchsteiger. 1998. “A Theory of Sequential Reciprocity.” Discussion Paper, CentER, Tilburg University.
- Falk, Armin, Ernst Fehr, and Urs Fischbacher. 1999. “On the Nature of Fair Behavior.” Working Paper. Zurich: University of Zürich, Institute for Empirical Research in Economics.
- Falk, Armin, and Urs Fischbacher. 1998. “A Theory of Reciprocity.” Working Paper. Zurich: University of Zürich, Institute for Empirical Research in Economics.
- Fehr, Ernst, Georg Kirchsteiger, and Arno Riedle. 1993. “Does Fairness Prevent Market Clearing? An Experimental Investigation.” *Quarterly Journal of Economics* 108(2):437-460.
- Fehr, Ernst and Klaus Schmidt. 1999. “A Theory of Fairness, Competition, and Cooperation.” *Quarterly Journal of Economics* 114:817-868.
- Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2000. “Are People Conditionally Cooperative? Evidence from a Public Goods Experiment.” Working Paper. Zürich: University of Zurich, Institute for Empirical Research in Economics.

- Forsythe, Robert, Joel L. Horowitz, and Nathan E. Savin. 1994. "Fairness in Simple Bargaining Games." *Games and Economic Behavior* 6(3):347-69.
- Franzen, Axel. 1994. "Group Size Effect in Social Dilemmas." In *Social Dilemmas and Cooperation*, ed. Ulrich Schultz, Wulf Albers, and Ulrich Mueller, 117-145. Berlin: Springer-Verlag.
- Friedman, James W. 1972. "A Non-cooperative Equilibrium for Supergames." *The Review of Economic Studies* 38:1-12.
- Friedman, Milton. 1954. *Essays in Positive Economics*. Chicago: University of Chicago Press.
- Fudenberg, Drew, and David M. Kreps. 1987. "Reputation in the Simultaneous Play of Multiple Opponents." *The Review of Economic Studies* 54:541-568.
- Fudenberg, Drew and David K. Levine. 1999. *The Theory of Learning in Games*. Cambridge, Mass.: The MIT Press.
- Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54:533-554.
- Fudenberg, Drew, and Jean Tirole. 1991. *Game Theory*. Cambridge, MA: MIT Press.
- Gneezy, Uri, Werner Güth, and Frank Verboven. 1998. "Presents of Investments? An Experimental Analysis." Working Paper. Berlin: Humboldt University.
- Goeree, Jacob K., Charles A. Holt, and Susan K. Laury. Forthcoming. "Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior." *Journal of*

Public Economics.

- Güth, Werner, and Hartmut Kliemt. 1998. "The Indirect Evolutionary Approach: Bridging the Gap Between Rationality and Adaptation." *Rationality and Society*. 10(3):377-399.
- Güth, Werner, Hartmut Kliemt, and Bezalel Peleg. 2000. "Co-evolution of Preferences and Information in Simple Games fo Trust." *German Economic Review* 1(1):83-110.
- Güth, Werner and M. Yaari. 1992. "An Evolutionary Approach to Explaining Reciprocal Behaviour in a Simple Strategic Game." in *Explaining Process and Change*, ed. Hartmut Kliemt, 23-34, Ann Arbor: University of Michigan Press.
- Hamburger, H., M Guyer, and J. Fox. 1975. "Group Size and Cooperation." *Journal of Conflict Resolution* 19(3):503-531.
- Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162 (December): 1243-1248.
- Hardin, Russell. 1982. *Collective Action*. Baltimore,MD: Johns Hopkins University Press.
- Harsanyi, J. C. 1955. "Cardinal Welfare, Individual Ethics, and Interpersonal Comaprisons of Utility." *Journal of Political Economy* 63:309-321.
- Harsanyi, J. C. 1967-68. "Games with Incomplete Information Played by Bayesian Players." *Management Science* 14:159-182, 320-334, 486-502.
- Hayashi, Nahoko, Elinor Ostrom, James Walker, and Toshio Yamagishi. 1999. "Reciprocity, Trust, and the Sense of Control: A Cross-Societal Study." *Rationality and*

- Society* 11(1):27-46.
- Hechter, M. 1992. "The Insufficiency of Game Theory for the Resolution of Real-World Collective Action Problems." *Rationality and Society* 4:33-40.
- Heckathorn, Douglas D. 1998. "Collective Action, Social Dilemmas and Ideology." *Rationality and Society* 10:451-479.
- Hobbes, Thomas. 1960[1651]. *Leviathan or the Matter, Forme and Power of a Commonwealth Ecclesiasticall and Civil*. Ed. Michael Oakeshott. Oxford: Basil Blackwell.
- Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith. 1996. "Social Distance and Other Regarding Behavior in Dictator games." *American Economic Review* 86(3): 653-660.
- Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith. 1998. "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology." *Economic Inquiry* 36(3): 335-352.
- Hogarth, Robin M. and Melvin W. Reder. 1986. "Editors' Comments: Perspectives from Economics and Psychology." *The Journal of Business* 59: S185 - S208.
- Hsiao, Cheng. 1996. "Logit and Probit Models" In *The Econometrics of Panel Data: Handbook of Theory and Applications*, ed. L. Matyas and P. Sevestre (eds.) Dordrecht, the Netherlands: Kluwer-Nijoff.
- Hurwicz, Leonid. 1973. "The Design of Mechanisms for Resource Allocation." *The American Economic Review* 63:1-30.

- Isaac, R. M., and James M. Walker. 1988 "Group Size Hypotheses of Public Goods Provision: The Voluntary Contribution Mechanism." *Quarterly Journal of Economics* 103(1):179-200.
- Isaac, R. M., James M. Walker, and Susan H. Thomas. 1984. "Divergent Evidence On Free Riding: An Experimental Examination of Possible Explanations." *Public Choice* 43:113-149.
- Jencks, Christopher. 1990. "Varieties of Altruism." In *Beyond Self-Interest*, ed. Jane J. Mansbridge, 53-60. Chicago: University of Chicago Press.
- Jones, Brooks, Matthew Steele, James Gahagan, and James Tedeschi. 1968. "Matrix Values and Cooperative Behavior in the Prisoner's Dilemma Game." *Journal of Personality and Social Psychology* 8:148-153.
- Kagel, John H., and Alvin E. Roth. eds. 1995. *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1986. "Fairness and the Assumptions of Economics." *Journal of Business* 59:S285-S300.
- Kelley, H. H., and J. W. Thibaut. 1978. *Interpersonal Relations*. New York: Wiley.
- Kollock, Peter. 1998. "Transforming Social Dilemmas: Group Identity and Cooperation." In *Modeling Morality, Rationality, and Evolution*, ed. Peter Danielson, 186-210. Oxford: Oxford University Press.
- Kramer, R. M., C. G. McClintock, and D. M. Messick. 1986. "Social Values and Cooper-

- ative Response to a Simulated Resource Conservation Crisis." *Journal of Personality* 54:576-592.
- Kreps, David M. 1997. "Intrinsic Motivations and Extrinsic Incentives." *American Economic Review* 87:359-364.
- Kreps, David M., P. Milgrom, J. Roberts, and R. Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma." *Journal of Economic Theory* 27:245-252.
- Kreps, David M., and Evan L. Porteus. 1979. "Dynamic Choice Theory and Dynamic Programming." *Econometrica* 47:91-100.
- Kreps, David M., and Robert Wilson. 1982a. "Reputation and Imperfect Information." *Journal of Economic Theory* 27:253-279.
- Kreps, David M., and Robert Wilson. 1982b. "Sequential Equilibria." *Econometrica* 50:863-894.
- Kuhlman, D. M., C. R. Camac, and D.A. Cunha. 1986. "Individual Differences in Social Orientation." In *Experimental Social Dilemmas*, ed. H. Wilke, D. M. Messick, and C. G. Rutte, 151-176. Frankfurt: Peter Land.
- Laffont, Jean-Jacque, and Michel Moreaux, eds. 1991. *Dynamics, Incomplete Information and Industrial Economics*. Translated by François Laisney. Cambridge, MA: Basil Blackwell.
- Latane, B., W. Kipling, and S. Harkins. 1979. "Many Hands Make Light the Work: The Causes and Consequences of Social Loafing." *Journal of Personality and Social*

- Psychology* 37(6):822-832.
- Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *The Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth, 111-194. Princeton, NJ: Princeton University Press.
- Luce, R. Duncan, and Howard Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.
- Mansbridge, Jane J. 1990. "The Rise and Fall of Self-Interest in the Explanation of Political Life." In *Beyond Self-Interest*, ed. Jane J. Mansbridge, 3-22. Chicago: University of Chicago Press. Pp.3-22.
- Marwell, Gerald, and Ruth E. Ames. 1979. "Experiments on the Provision of Public Goods I: Resources, Interest, Group Size, and the Free Rider Problem." *American Journal of Sociology* 84:1335-1360.
- Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge, UK: Cambridge University Press.
- McCabe, Kevin A., and Vernon L. Smith. 1999. "Strategic Analysis by Players in Games: What Information Do They Use?" Working Paper. Tucson: University of Arizona.
- McCabe, Kevin A., Vernon L. Smith, and Michael LePore. 1998. "Intentionality Detection and 'Mindreading': Why Does Game Form Matter?" Working Paper. Tucson: University of Arizona.
- Mertens, Jean-Francois, and S. Zamir. 1985. "Formulation of Bayesian Analysis for Games

- with Incomplete Information." *International Journal of Game Theory* 14:1-29.
- Miller, Gary J. "The Impact of Economics on Contemporary Political Science." *Journal of Economic Literature* 35:1173-1204.
- Moreaux, Michel. 1985. "Perfect Nash Equilibria in Finitely Repeated Game and Uniqueness of Nash Equilibrium in the Constituent Game." *Economics Letters* 17:317-320.
- Nash, John F. 1950. "Equilibrium Points in N-person Games." *Proceedings of the National Academy of Sciences, U.S.A.* 36:48-49.
- Ochs, Jack, and Alvin E. Roth. 1989. "An Experimental Study of Sequential Bargaining." *American Economic Review* 79(3):355-384.
- Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Group*. Cambridge, MA: Harvard University Press.
- Orbell, John and Robyn M. Dawes. 1991. "A 'Cognitive Miser' Theory of Cooperators' Advantage." *American Political Science Review* 85:515-528.
- Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.
- Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action." Presidential Address, American Political Science Association, 1997. *American Political Science Review* 92:1-98.
- Ostrom, Elinor. 2000. "Collective Action and the Evolution of Social Norms." *Journal of Economic Perspectives* 14(3):137-158.

- Ostrom, Elinor, Roy Gardner, and James M. Walker. 1994. *Rules, Games, and Common-Pool Resources*. Ann Arbor: University of Michigan Press.
- Ostrom, Elinor, and James M. Walker. 1997. "Neither Markets Nor States: Linking Transformation Processes in Collective Action Arenas." In *Perspectives on Public Choice: A Handbook*, ed. Dennis C. Mueller, 35-72. Cambridge: Cambridge University Press.
- Ostrom, Vincent, and Timothy Hennessey. 1975. "Institutional Arrangements in a Market Economy." In *Conjectures on Institutional Analysis and Design: An Inquiry into Principles of Human Governance*, Drafted manuscript, Workshop in Political Theory and Policy Analysis, Indiana University, Bloomington.
- Palfrey, T. R., and J. E. Prisbrey. "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *American Economic Review* 87:829-846.
- Rabin, Matthew. 1993. "Incorporating Fairness in Game Theory and Economics." *American Economic Review* 83(5):1281-1302.
- Rapoport, Anatol. 1988. "Experiments with N-Person Social Traps I and II." *Journal of Conflict Resolution* 32:457-488.
- Rapoport, Anatol, and A. M. Chammah. 1965. *Prisoner's Dilemma*. Ann Arbor: University of Michigan Press.
- Rasmusen, Eric. 1989. *Games and Information*. Cambridge, MA: Blackwell Publishers.
- Raub, Werner. 1990. "A General Game-Theoretic Model of Preference Adaptations in Problematic Social Situations." *Rationality and Society* 2:67-93

- Roth, Alvin E. 1995. "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth, 3-109. Princeton, NJ: Princeton University Press.
- Roth, Alvin E. 1991. "Game Theory as a Part of Empirical Economics." *The Economic Journal* 101:107-114.
- Saijo, T., and H. Nakamura. 1995. "The 'Spite' Dilemma in Voluntary Contributions Mechanism Experiments." *Journal of Conflict Resolution* 39:535-560.
- Satz, Debra, and John Ferejohn. 1994. "Rational Choice and Social Theory." *The Journal of Philosophy* 91:71-87.
- Schmidt, David, Robert Shupp, James M. Walker, Elinor. Ostrom, and T.K. Ahn. Forthcoming. "Dilemma Games: Game Parameters and Matching Protocols." *Journal of Economic Behavior and Organization*.
- Schumpeter, Joseph A. 1942. *Capitalism, Socialism, and Democracy*. New York and London: Harper and Brothers Publishers.
- Sefton, Martin, Robert Shupp, and James Walker. 2000. "The Effect of Rewards and Sanctions in Provision of Public Goods." Working Paper. Bloomington: Indiana University.
- Selten, Reinhard. 1991. "Evolution, Learning, and Economic Behavior." *Games and Economic Behavior* 3:3-24.
- Selten, Reinhard and Rolf Stoecker. 1986. "End Behavior in Sequences of Finite Prisoner's

- Dilemma Supergames: A Learning Theory Approach." *Journal of Economic Behavior and Organization* 7:47-70.
- Sen, Amartya K. 1974. "Choice, Orderings and Morality." In *Practical Reason: Papers and Discussions*, ed. Stephan Körner, 54-82. Oxford: Blackwell.
- Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economics Theory." *Philosophy & Public Affairs* (Summer): 317-344.
- Sen, Amartya K. 1985. "Well-Being, Agency and Freedom: The Dewey Lectures 1984." *The Journal of Philosophy* 82:169-221.
- Shafir, Eldar, and Amos Tversky. 1992. "Thinking through Uncertainty: Non-consequential Reason and Choice." *Cognitive Psychology* 24:449-474.
- Silveria, Rava da. 1999. "An Introduction to Breakdown Phenomena in Disordered System." *American Journal of Physics* 62:1177-1188.
- Skinner, B. Frederic. 1974. *About Behaviorism*. New York: Knopf.
- Smale, S. 1980. "The Prisoner's Dilemma and Dynamic Systems Associated to Non-Cooperative Games." *Econometrica* 48:1617-1634.
- Smith, Vernon L. 1982. "Microeconomic Systems as an Experimental Science." *American Economic Review* 72:923-955.
- Smith, Vernon L. 1991. *Papers in Experimental Economics*. New York: Cambridge University Press.

- Smith, Vernon L. 1997. "The Two Races of Adam Smith." Southern Economic Association Distinguished Guest Lecture.
- Taylor, Michael. 1987. *The Possibility of Cooperation*. New York: Cambridge University Press.
- Tullock, Gordon. 1992. "Games and Preference." *Rationality and Society* 4:24-32.
- Van Kolpin, W. 1993. "Shared Facility Games with Variable Utilization." *International Economic Review* 34(2):387-400.
- Von Neumann, John, and Oskar Morgenstern. 1953[1944]. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
- Wagner, Thomas. 1998. "Reciprocity and Efficiency" *Rationality and Society* 10(3):347-376.
- Walker, James, Roy Gardner, and Elinor Ostrom. 1990. "Rent Dissipation in a Limited-Access Common-Pool Resource: Experimental Evidence." *Journal of Environmental Economics and Management* 19:203-211.
- Weibull, Jörgen W. 1998. "Evolution, Rationality and Equilibrium in Games." *European Economic Review* 42:641-649
- Wildavsky, A. 1992. "Indispensable Framework or Just Another Ideology? Prisoner's Dilemma as an Antihierarchical Game." *Rationality and Society* 4:8-23.
- Yamagishi, Toshio. 1997. "Everyone Is a Thief!?: Generalized Distrust and Social Intelligence." Paper prepared for presentation at the International Conference of Distrust,

Bellagio, Italy, October 14-16, 1997.

Yamagishi, Toshio, and Karen S. Cook. 1993. "Generalized Exchange and Social Dilemmas." *Social Psychology Quarterly* 56(4):235-248.

Yamagishi, Toshio, and Tohko Kiyonari. 1998. "Playing a Prisoner's Dilemma as an Assurance Game: Matrix Transformation and Production of Trust." Paper prepared for the Trust Conference, New York, November 14-16.

Toh-Kyeong Ahn

Postdoctoral Fellow
Workshop In Political Theory and Policy Analysis
Indiana University
513 North Park Avenue
Bloomington, IN 47408
Phone (812) 855-0441
Fax (812) 855-3150
E-mail: tahn@indiana.edu

Education:

- Ph. D., Political Science, 2001, Indiana University
Subfields: Theory and Methodology (Institutional Analysis/ Game Theory/Statistics)
Public Policy (Public Choice/Policy Analysis)
Economics (Minor)
Dissertation: "Foundations for Cooperation in Social Dilemmas"
Advisors: Elinor Ostrom (Chair), Mike McGinnis, Burt Monroe
James Walker, Pravin Trivedi
- M.A., Political Science, 1994, Seoul National University
- B.A., Political Science, 1990, Seoul National University

Publications:

Journal Articles

- Ahn, T.K., E. Ostrom, D. Schmidt, R. Shupp, and J. Walker. 2001. "Cooperation in PD Games: Fear, Greed, and History of Play." *Public Choice* 106(1/2):137-155.
- Gibson, Clark, Elinor Ostrom, and T.K. Ahn. 2000. "The Concept of Scale and the Human Dimensions of Global Environmental Change." *Ecological Economics* 32: 217-239.
- Schmidt, D., R. Shupp, J. Walker, E. Ostrom, and T.K. Ahn. Forthcoming. "Dilemma Games: Game Parameters and Matching Protocols." *Journal of Economic Behavior and Organization*.

Book Chapter

- Ahn, T.K., E. Ostrom, D.Schmidt, and J. Walker. forthcoming. "Trust in Two Person Games: Game Structure and Linkage." In *Trust, Reciprocity, and Gains from Association: Interdisciplinary Lessons from Experimental Research*, ed. Elinor Ostrom and James Walker. New York: Russell Sage Foundation.

Monograph:

- Gibson, Clark, Elinor Ostrom, and T.K. Ahn. 1998. "Scaling Issues in the Social Sciences: A Report for the International Human Dimensions Programme on Global Environmental Change." International Human Dimensions Programme Working Paper No. 1, Bonn, Germany.

Working Papers:

- "A Social Science Perspective on Social Capital: Social Capital and Collective Action." (with Elinor Ostrom). Indiana University, Workshop in Political Theory and Policy Analysis, Working Paper #W01-2.
- "Finitely Repeated 2x2 Social Dilemma Games: Equilibrium Analysis and Experimental Test." Indiana University, Workshop in Political Theory and Policy Analysis, Working Paper #W00-25.
- "Group Size and Collective Action: The Impossibility of Payoff Parameters Control in N-Person Prisoner's Dilemma." Indiana University, Workshop in Political Theory and Policy Analysis, Working Paper #W00-26.
- "Altruism or Equity?: Modeling Non-Selfishness in Social Dilemmas." (with Elinor Ostrom and James Walker). Indiana University, Workshop in Political Theory and Policy Analysis, Working Paper #W98-34.
- "Communication and Cooperation in Common-Pool Resources Dilemma Experiments: A Test of Experimental Validity." (with Elinor Ostrom) Indiana University, Workshop in Political Theory and Policy Analysis, Working Paper.

Awards and Fellowship:

- Dissertation Fellowship (1999-2000), Workshop in Political Theory and Policy Analysis, Indiana University.

Academic and Teaching Experiences:

- Arthur Bentley Research Assistantship (Fall, 1998), Department of Political Science, Indiana University
- Research Assistantship (Fall, 1996- Spring, 1997), Workshop In Political Theory and Policy Analysis, Indiana University
- Associate Instructor, Department of Political Science, Indiana University
 - Y575: Political Data Analysis I (Fall, 1997)
 - Y576: Political Data Analysis II (Spring, 1999)
 - Y572: Mathematical Tools for Political Scientists (Spring, 1999)