&Part III Game Theory and the Institutions Within Ourselves: Social Propensities

&Chapter 8 We Are Friends, Right? Social Relationships

8.1 Embeddedness and Face-to-Face Interactions
8.2 The Stimulus-Response Mechanism
8.3 The Economic Value of Friendship
8.4 Who Needs You Anyway? Social Ostracism
8.5 Playing Social Games

&Chapter 9 We Have Feelings Too! Social Preferences

9.1 Social Preferences: Altruism, Inequality, and Social Welfare
9.2 Social Preferences: Reciprocity and Caring About Others' Intentions
9.3 Rational Behavior and Social Preferences
9.4 Looking Ahead

&Chapter 10 Social Preferences, Norms, Emotions, and Internalized Institutional Elements

10.1 Social Preferences, Rules, and Behavioral Cultural Beliefs: A Dispositional Perspective
10.2 Incomplete Information and Organizations
10.3 The Contingency of Social Preferences: A Situational Perspective
10.4 Social Preferences and Internalized Institutional Elements
10.4.1 Studying Internalized Norms
10.4.2 Studying Emotions using Psychological Games
10.5 Rules and Communication
10.6 The Evolutionary Origin of Prosocial Preferences
10.7 Looking Ahead

Adam Smith is best known in economics for his assertion in the *Wealth of Nations* (1776: 13) that markets are a "consequence of a certain propensity in human nature ... the propensity to truck, barter, and exchange one thing for another." Yet, Adam Smith was keenly aware of other, more subtle ways though which other human propensities are crucial to the operation of markets and economic outcomes. In his earlier work, *Theory of Moral Sentiments* (1759), he examined, for example, how one's desire for sympathy facilitates exchange and the functioning of markets. One's desire for sympathy from others transforms his self-interest into an adherence to promises. Hence, the desire for sympathy facilitates exchange by promoting the expectation that promises will be kept. Various human propensities can imply non-material considerations that can inter-relate with a society's institutions and influence behavior.

Sociologists paid heed to arguments regarding how various human propensities inter-relate with a society's institutions. They have maintained that studying institutions first requires examining their inter-relations with human propensities, such as the desire for sympathy by one's peers, the ability to integrate values, to care for others, and sensitivity to other's intentions. Parsons (1951: 38-40) has taken the position that full institutionalization of a behavioral standard requires its internalization, while Durkheim (1953: 129) has similarly argued that institutions are "something beyond us and something in ourselves." As we have seen, "old institutionalism" held a similar position.

Many prominent economists, such as Akerlof 1986, Arrow 1981, Hirshleifer 1985, Becker 1974, Lal 1988, North 1990, Platteau 1994, Samuelson 1993, and Sen 1993, have argued that such considerations should be integrated into the field of economics.[1] Nevertheless, until recently they remained a marginal area of exploration in economics. This has been the case, to a large extent, because of the lack of an appropriate analytical framework. After all, by arbitrarily defining such unobservable non-material factors as peer pressure, values, duty, or fairness, we can account for any observed behavior.

In game-theoretic terms, the problem can be illustrated using the following a prisoners'

[1] Indeed, such assertions were recently substantiated through cross-country regressions. See with respect to trust, for example, Knack and Keefer 1997, Glaeser, et al. 2000.

.

dilemma (PD) game.

	Cooperate	Cheat
Cooperate	1, 1	-1, 2
Cheat	2, -1	0, 0

In this game, the only self-enforcing behavior is for both players is to cheat. Each player's payoff is higher if he or she cheats independently of the other player's action. (Section 4.1) If this game captures all aspects of the situation, the only beliefs that can prevail are that both players will cheat.

Now suppose that the one also has preferences over his own actions. Playing "cooperate" also implies satisfaction from "fulfilling one's duty" while playing cheat implies dissatisfaction from taking the "immoral" action. To capture the implications of such non-material rewards and sanctions and their relationships with the action of the other player, consider an augmented version of the PD game. This game also incorporates non-material payoffs in addition to the above material payoffs. Cooperating is "doing the right thing," implying the added non-material payoff of α. Cheating, when the other cooperates is "acting inappropriately," entailing a non-material cost of γ. ($\alpha,\gamma \geq 0$.) The normal form of this augmented game is:

	Cooperate	Cheat
Cooperate	$1 + \alpha$, $1 + \alpha$	$-1 + \alpha$, $2 - \gamma$
Cheat	$2 - \gamma$, $-1 + \alpha$	0 , 0

In this augmented game, the belief that both players will cooperate can prevail. This is the case if the combined utility derived from cooperation if the other cooperates is higher than the combined utility from cheating when the other cooperates, that is, $-1 + \alpha > 2 - \gamma$. If both players maintain that this condition holds, mutually beneficial exchange will be undertaken in a situation in which it otherwise would not have been possible. Furthermore, if $\alpha \geq 1$, this

behavior is the only self-enforcing behavior.

To the extent that such non-material factors as those captured by α and γ are outcomes of human actions, and are non-technological factors exogenous to each of the individuals whose behavior they influence, they are institutional elements. They differ from the behavioral cultural beliefs considered in part II in providing motivation without relying on inflicting or providing material sanctions or rewards. Such institutional elements build on various humans' social propensities such as sensitivity to others' opinions, enjoyment derived from social interactions, statues and respect, and the ability to internalize norms of behavior. Accordingly, it seems appropriate to refer to these institutional elements as internalized institutional elements.

The deficiency of the above specification for studying internalized institutional elements is that it is arbitrary. Examining internalized institutional elements in a way that is consistent with economic methodology requires clear specification of the issues under consideration and an appropriate analytical framework. An analytical framework is one which deductively restricts admissible arguments regarding internalized institutional elements that can prevail in a given environment; enables studying the inter-relationships between internalized and other institutional elements; and makes explicit how internalized institutional elements influence behavior and are generated by it.

Game theory, augmented by insights from various disciplines other than economics, made various contributions toward developing such an analytical framework and this part elaborates on these contributions. It presents the usefulness of game theory for exploring issues central to old institutionalism and the sociological approach to institutions. Before presenting these contributions, it should be recalled that this part, similar to part II, neither examines the origin of institutions nor their changes. Its focus is on studying institutions as a steady-state equilibrium in which institutional elements generate behavior which, in turn, generates these institutional elements. It particularly concentrates on the game-theoretic contributions to identify and model various relevant human propensities and their manifestations, deductively restricting admissible arguments regarding internalized institutional elements, understanding how particular internalized institutional elements can generate behavior and be generated by it, and how they inter-relate with behavior and other institutional elements.

This part is organized as follow. Chapter 8 concentrates on social relationships and

considers the implications of two human propensities. The first is sensitivity to the feeling of those with whom one interacts face-to-face. The second is the human tendency to value social interactions for their own sake, that is, to gain enjoyment from social interactions. The chapter concentrates on how game theory provides an analytical framework to study the implications of these social propensities.

Chapter 9 presents the contributions of game-theoretic experiments expose manifestations of humans' capacity to have social preferences: to be sensitive to the welfare of others and the intentions behind their actions. The framework provided by game theory to conduct experiments revealing social preferences enabled exact formulation of their manifestations. The chapter also presents the insights from these experiments on the extent to which individuals act strategically.

Chapter 10 explores the implications of social preferences for institutional analysis. It highlights two main approaches. The first is closely related to the approach presented in parts I and II. In these parts we examined institutions assuming that individuals are selfish. The same analysis is made here while replacing this assumption with the assumption that some individuals have social preferences. The second approach recognizes that the manifestations of the human capacity to have social preferences are socially determined. These manifestations reflect the social malleability of preferences. This view implies the need to study social preferences within the broader context of the social construction of internalized norms and emotions. The chapter thus elaborates on the game-theoretic contributions to studying these issues. It should be noted that the analysis elaborated in this part is still tentative. Accordingly, I do not provide any extensive empirical example.

.

&Chapter 8 We Are Friends, Right? Social Relationships

The above reference to Adam Smith in the *Theory of Moral Sentiments*, reflects an assertion about an important human propensity to seek social approval and to avoid social sanctions and disapproval. Sociologists have long argued that this propensity plays an important role in institutions, leading to social order and cooperation. This argument has been made as early as Bernhard de Mandeville's (1714) *Fable of the Bees,* while in modern sociology it is particularly associated with the George Homans (1961) and Dennis Wrong (1961). Dennis Wrong (1961) has nicely summarized this sociological perspective: "it is frequently the task of the sociologist to call attention to the intensity with which men desire and strive for the good opinion of their immediate associates in a variety of situations."[2]

Such considerations and how they contribute to motivating individuals to take particular actions cannot be captured in, for example, the neo-classical framework. In it individuals are assumed to interact anonymously and only in the economic arena. Game theory, however, enables modeling interactions among specific individuals and provides an analytical framework within which one can examine the implications of multi-dimensional interactions among them. In other words, this analytical framework enables examining the implications of having the same individuals interacting socially and, for example, economically, inside or outside a larger group of people engaged in similar interactions. Game theory exposes the conditions under which particular economic outcomes are possible given such multi-dimensional interactions.

Hence, game theory further contributes to the study of social norms, as defined in section 7.1. Rules that are neither promulgated by an official source, such as a court or a legislature, nor enforced by the threat of legal sanctions, yet are regularly complied with. Unlike the discussion in part II, however, the motivation to comply is based here on social, rather than material, incentives.

The discussion is organized around the two main ways in which game-theoretical analysis has been incorporated with the study of social relationships. Not surprisingly, these two ways differ in how they treat the need to ensure the credibility of punishments and rewards. The

---

[2] Cited in Granovetter 1985: 483. Wong has qualifications regarding this position as discussed below.

first approach adopted the psychological stimulus-response mechanism in which the credibility of the threat is based on emotions. (Sections 8.2 and 8.3.) In the second approach the credibility of social ostracism is achieved based on either the limited contribution of any particular individual to the group or the expectation that failure to punish a deviator would lead others to deviate as well. (Section 8.4.) In addition the chapter elaborates on the distinction between economic and social relationships (section 8.1) and Homans's (1950) theory of endogenous social affection (section 8.3).

To simplify the presentation, the following discussion takes the information structure of various games as given. People are assumed to know, for example, who took what action in the past. Clearly, as has been emphasized in part II, the transmission of information has to be endogenous to the analysis.

8.1 Embeddedness and Face-to-Face Interactions

Perhaps the formulation of the importance of social relationships best known to economists is associated with the work of Granovetter (1985) on embeddedness.[3] Granovetter argues that economic behavior is always embedded in ongoing social relations. "Actors do not behave or decide as atoms outside a social context, nor do they adhere slavishly to a script written for them by the particular intersection of social categories that they happen to occupy. Their attempts at purposive action are instead embedded in concrete, ongoing systems of social relations" which may be bilateral or within a social structure, a network (p. 487).

Why are ongoing social relations important? They are important for the same two reasons that social relationships were important among the Maghribi traders. They change the information structure of interactions and they facilitate enforcement, motivating individuals to take particular actions. An important aspect of Granovetter's argument is exactly how social interactions influence motivation. He argues that "continuing economic relationships often become overlaid with social content that carries strong expectations of trust and abstention from opportunism" (460). Personal, face-to-face economic interactions create a social bond among the interacting individuals which arguably motivate them to forego economically rewarding

[3] For another formulation, see that of "cojointness" in Coleman 1990.

opportunistic behavior. Face-to-face, personal interactions generate the expectation of trust and trust itself.

Retrospection and experiments confirms these assertions, even in the absence of continuing social relationships. Asch's (1952) classic experiment showed the power of groups to generate conformity. Asch recruited students allegedly for a study of visual perception. Before the experiment began, he explained to all but one member in a small group that their real purpose was to put pressure on the remaining person. Arranging six to eight students around a table, Asch showed them a "standard" line on one card and asked them to match it to one of three lines shown on another card. Anyone with normal vision could easily see what the correct match was. Initially, as planned, everyone made the matches correctly. But then Asch's secret accomplices began answering incorrectly, leaving the naive subject (seated at the table in order to answer next-to-last) bewildered and uncomfortable. What happened? Asch found that one-third of all subjects conformed to the others by answering incorrectly. Many are apparently willing to compromise their own judgment to avoid being different, even from people they do not know.

More recent experiments have similarly revealed that people are willing to sacrifice monetary income to do what is good for a small group.[4] Furthermore, experiments indicate that face-to-face communication, the ability to talk with one another, has a profound effect in fostering cooperation. In games in which one's best response is not to cooperate by taking an action, such as playing defect in a prisoners' dilemma game or not contributing to the public good, face-to-face communication led to a sharp rise in cooperation (Ostrom 1998: 6-7).

Perhaps the best evidence for the importance of face-to-face interactions in influencing behavior effecting material outcomes comes from experiments in the Dictator Game developed by Forsythe, et al. (1994). In this game, one individual, the dictator, could divide m dollars between himself and someone else. It was found that when individuals had to make this choice and divide $10, only 18 percent offered $0 and 32 percent offered $4 or more. Hoffman et al. 1994, repeated this experiment using a double-blind procedure intended to guarantee the complete social isolation of the individual's decision and this was known to the participants. No one, including the experimenter or any subsequent observer of the data, could possibly know any

[4] Dawes and Thaler 1988: 194-5.

subject's decision. In this case, 64 percent of the offers were $0 with only 8 percent offering $4 or more. The difference between the experiments is statistically significant. (See similar results in Hoffman et al. 1996). Bohnet and Frey (1999) found that in the Dictator Game with full anonymity, only 25 percent of the dictators choose equal division. This amount increased to 71 percent when the two players were identified to each other.[5]

8.2 The Stimulus-Response Mechanism

Completing the argument regarding the importance of face-to-face interactions requires examining how it inter-relates exactly with motivation and economic behavior. Psychologists have argued that face-to-face interactions provide motivation through the "stimulus-response" mechanism. Humans are emotionally predisposed to be sensitive to what they think others feel about them. One is sensitive to what he believes others, whom he knows, think of him. Such social approval and disapproval is basically emotional and uncontrollable and may indeed exist only in the mind of the recipient. (Homans 1961 and Smith 1759.) One feels shame from the perception that a waiter who did not get an expected tip is resentful even if the waiter doesn't show it. The expectation or perception of others' emotionally prompted responses influence the overall utility that one derives from taking a particular action. Under the stimulus-response mechanism, the provision of rewards and sanctions is credible because they are not under the control of the one who provides or inflicts them.

To accommodate this mechanism in a game-theoretic framework it is necessary to take these steps: first, the payoff specification should include, in addition to one's material payoffs from various actions, one's payoff from what he perceives others think of him; second, the origin of this perception. One way to do so is to specify that one's perception about others' thoughts of him reflect the difference between his behavior and the behavior of others. Once these changes in the basic game structure are made. An equilibrium analysis restricts the set of outcomes that can prevail in the environment under consideration while capturing the combined influence of

[5] Hoffman et. al. argue that their findings support the view that people bring their everyday rules-of-thumb concerning repeated interactions to a single-shot game. Bohnet and Frey argued, however, that in face-to-face interactions increased equality of the division reflect the fact that the other is now an "identifiable victim." This difference is not important to the argument made here.

material and non-material considerations on one's behavior.

A wonderful analysis of this sort has been conducted by Holländer (1990) who has employed a game-theoretic model (which is too complex to be presented here) to examine voluntary provision of public good. In public good provision games, economic considerations provide no incentive to contribute to the provision of public good. Each individual prefers that others contribute to the provision of public good, implying that no one does.

This is no longer the case when one assumes that individuals are responsive to emotionally prompted social approval by individuals in the relevant reference group. One's desire to gain social approval from other members of society influences his economic behavior. Although the extent to which this desire influences behavior is exogenous to an individual, it is endogenous to the group of interacting individuals. Each individual in choosing behavior considers the economic cost of contributing a particular amount to the public good and the social approval and disapproval it implies. The social approval or disapproval that a particular action implies, in turn, is proportional to the actions other individuals have taken. Examining this situation as a game yields the equilibrium outcome of contribution to the public good as well as the extent of social, emotionally motivated approval and disapproval.

??? A simple version of the model will be provided to illustrate its contribution to deductively restricting assertions.

8.3 The Economic Value of Friendship

Holländer's analysis took for granted that individuals care about what other think of them. The sociologist, Homans, however, considers the conditions that would lead to such a situation. Specifically, he proposed a model capturing how "friendship" can endogenously emerge among individuals. His formulation can be integrated in a game-theoretic framework that captures its implications on the ability to resolve commitment problems.

Homans (1950) has argued that the amount of activity and friendship among members of a group reflect the dynamic interactions among four variables: A(t) - the amount of activity carried on by the members within the group; I(t) - the intensity of interactions among the members; F(t) - the level of friendliness among the members; E(t) - the amount of activity

imposed on the group by the external environment.

To illustrate the relationship between activity and the intensity of interaction consider the case of "two men at opposite ends of a saw, sawing a log. When we say that the two are interacting, we are not referring to the fact that both are sawing: in our language, sawing is an activity, but to the fact that the push of one man on the saw is followed by the push of the other. In this example, the interaction does not involve words. More often interaction takes place through verbal or other symbolic communication" (Homans 1950: 36). Friendliness is a particular sentiment that may be thought of as an indicator of the level of utility that an individual derives from his concrete social relations with other group members.[6]

Homans develops a theory of the dynamic relationships among these variables whose essence can be presented by a system of equations.[7] These variables reflect the behavior of an individual within the group, and should be considered to represent averages. (a) The intensity of interaction increases with the level of friendliness and activity carried on within the group. That is $I(t) = a_1F(t) + a_2A(t)$ were I(t), for example, represents the level of I at time t. (b) The level of friendliness will increase if the actual level of interaction is higher than "appropriate" to the existing level of friendliness.[8] (c) The level of activity carried on by the group will tend to increase if the actual level of friendliness is higher than "appropriate" to the existing amount of activity, and if the amount of activity imposed externally on the group is higher than the existing amount of activity.[9]

A steady state equilibrium in this system is one in which the variables remain stationary. Solving for it yields the following relationships between the exogenous variable, E_0, and the

---

[6] See discussion in Homans 1950: 37 ff., 115.

[7] This presentation draws on Simon 1987: 100 ff. The assumptions of linear relationships and that all adjustments are instantaneous, are made without loss of generality for ease of exposition. This presentation assumes that each of these variables can be represented, each moment of time, in terms of real numbers. Clearly, since the units in which such variables can be measured are arbitrary, only ordinal investigation is meaningful.

[8] That is: $dF(t)/dt = 1(t) - a_3F(t)$.

[9] That is: $dA(t)/dt = [F(t) - a_4A(t)] + [E(t) - A(t)]$.

endogenous variables.[10] While the various parameters of the system have meaning that can facilitate empirical analysis, to make the benefit of this model transparent, it is sufficient to note that for given amount of externally imposed level of activity, E_0, the equilibrium is characterized by: $A^* = (\cdot, E^0)$ and $F^* = F(\cdot, E^0)$ where both F^* and A^* increase in E^0.[11] In particular, the higher the level of external activity imposed on the group, E^0, the higher the amount of voluntary social activity, A^* ,and friendship, F^*, that will be generated. After sawing logs for several hours, it would not be surprising if the two woodcutters ended up drinking or chatting together. The social and the economic activities individuals engage in generate a level of sentimental relations and of friendship among them.

The level of friendship here can be interpreted as capturing one's level of sensitivity to the approval or disapproval of one's actions by others based on the psychological stimulus-response mechanism. One is sensitive to hurting the feelings of a friend. Expecting that, one can trust an acquaintance or a friend more than a stranger. Economic cooperation between specific individuals can be self-enforcing in situations in which it would otherwise not be forthcoming

We can show this implication of personal relationships in a game-theoretic model that captures the idea that one trusts a friend more. Consider, for example, figure 8.1 which is a social version of an exchange game. In this game one's payoff depends both on the actions taken and the level of friendship, where the level of friendship can be thought of as emerging

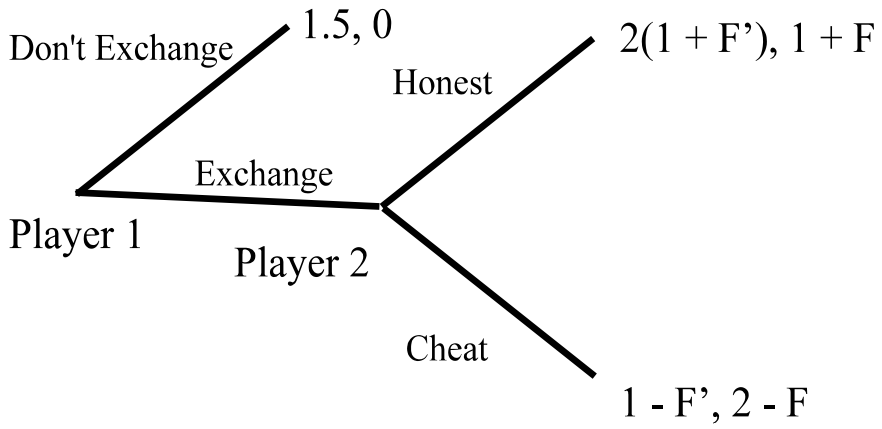
---

[10] An equilibrium is characterized by: $I^* = a_1F^* + a_2A^*$, $I^* - a_3F^* = 0$, $[F^* - a_4A^*] + [E^*-A^*] = 0$. Solving this system for A^*, F^* yields (where $g = (a_4 + 1)(a_3 - a_1) - a_2$). $A^* = a_3 - a_1/gE_0$ and $F^* = a_2/gE_0$. The stability conditions for this equilibrium are: $g > 0$ and $a_3 > a_1$. From the equilibrium stability value, the following comparative static results can be obtained: $dF^*/da_2 = (a_4 +1)(a_3 - a_1)/g^2 > 0$; $dA^*/da_2 = (a_3 - a_1)/g^2 > 0$; $dF^*/da_1 = a_2(a_4 + 1)/g^2 > 0$; $dA^*/da_1 = dF^*/da_4 = dA^*/da_3 = -a_2/g^2 < 0$; $dF^*/da_3 = -(a_4 +1)a_2/g^2 < 0$; $dA^*/da_4 = -(a_3 - a_1)^2/g^2 < 0$.

[11] The various coefficients of this system have intuitive meaning. In particular, a1 and a2 measure the amount of interaction generated per unit of friendliness and activity respectively. Thus, they may be called the "coefficient of interdependence." The variable a3F is the amount of interaction "appropriate" to the level F of friendliness. Thus, the reciprocal of a3 might be called the "congeniality coefficient" since it measures the amount of friendliness that will be generated per unit of interaction. The reciprocal of a4 measures the amount of activity generated "spontaneously" per unit of friendliness. Thus, it may be called the coefficient of "spontaneity." In equilibrium, $A^* = A(a_1, a_2, -a_3, -a_4, E^0)$ and $F^* = F(a_1, a_2, -a_3, -a_4, E^0)$ where both F^* and A^* increase in each of their elements

from interactions in the system described above. The payoff captures material (economic) factors but also non-material factors, specifically, that one trusts his friend, one is disappointed by a friend's betrayal, and disappointing a friend is psychologically costly.

Figure 8.1. Social Exchange Game: Friendship.



In the above game the payoffs for both players from honest behavior increases in the level of their friendship but decreases in this level if cheating occurs. The level of friendship is normalized to be between 0 and 1. Hence, if prior to playing the game players are engaged in some externally imposed activity leading to a positive and sufficiently high level of friendship, it can enable cooperation in the exchange game. Specifically, if $1 + F \geq 2 - F$, player 2's best response if exchange was initiated is to be honest. If player 1 anticipates this, he can trust player 2 and initiate exchange. Indeed, the only sub-game perfect equilibrium entails exchange.

Combining a model of friendship formation, the stimulus-response mechanism, and a game- theoretic model of an individual's decision-making make it possible to explore the inter-relationships between the environment, social considerations, and material outcomes. Such a model can provide, for example, a complete account for Homans's (1950: 334 ff.) classical study of the history of the New England city of Hiltown.

Early in the twentieth century, given the nature of the economic system and available

transportation, most of the residents of Hiltown worked within or around the town, shopped in it, and social life flourished in the city's public places and private homes. The external conditions imposed a high level of activity among the residents. In other words, E^0 was high. The equilibrium level of friendship, F*, was thus presumably high as well, and a dense network of information transmission and "gossip" prevailed, implying that everyone would have known what actions other people took in public situations.

Such relevant public situations were those associated with the provision of a public good. High levels of friendship and the stimulus-response mechanism imply that in a game of public good provision, individuals will contribute a lot. The free-rider problem - everyone prefers that someone else provide the public good - could have been mitigated by people's desire for social approval. Indeed, the city recorded a high score in its ability to overcome collective action problems and benefitted from a high level of public good provision. Groups of neighbors (the "bees") collaborated in particularly difficult tasks. Each of the three churches supported a women's organization, a young people's club, and a Sunday School. The Unitarians formed a society, the Social Union, that included both men and women, and held biweekly socials. These groups carried out works of charity and raised money for the church. The town meetings were well attended.

Changes in transportation technology that caused the town folks to became integrated into the wider economy and society changed this situation. By 1945 most of its residents worked and shopped mainly outside the town, their social activities were mostly out of it and mainly entailed going to the theater or restaurants in near-by larger towns. In other words, the level of exogenously imposed activities among the residents, E_0, substantially declined. According to the above model this reduction should have led to a lower level of equilibrium friendship, F^*, and limited the ability to overcome the free-rider problem associated with the provision of a public good. Indeed, the supply of public good provided voluntarily by members of the community drastically declined. Only one young people's club remained active, charitable activities declined, and the town's public meetings were attended by only a few.

In the above example, exogenous factors - transportation technology and the nature of the economy - led to the initial activities and interactions that led to friendship that sustained further economic activities by providing the bond required for commitment. The initial activities and

interactions that can lead to such friendship can also emerge endogenously. For example, we have already seen that one may be able to commit to honesty in exchange based on fearing the loss of future gains from future economic exchange. Such (economically-based) reputation mechanisms can constitute an activity, leading to interactions and friendship that can, in turn, provide the enforcement required for exchange that otherwise would not be self-enforcing A virtuous cycle of cooperation, friendship, and further cooperation can thus evolve. Game theory provides a framework within which we can capture this virtuous cycle.

Consider a repeated version of the exchange game explored in chapter 4. In this game people exchange based on a purely economic motivation: player 1 initiates exchange because player 2's best response is to be honest, currently fearing that cheating will entail losing gains from future exchange. But in light of the above discussion, we can extend this game to a social one. Once exchange has began and the two individuals repeatedly interact over time, friendship evolves, creating a social bond between them. The economic game has been transformed into a social game which, in turn, enables economic cooperation that would not have been possible otherwise. To illustrate this point, consider the following Social Prisoners' Dilemma game (SPD).

In each period of this infinitely repeated game, the players face a continuum of PD games in each of which a player can either Cheat (C) or Defect (D). Playing (D,D) in a particular PD game is interpreted as the parties choosing not to interact in that game. For a game to be played in a particular period both players have to choose to do so. The games differ in the amount that can be gained from cooperation. Specifically, in each game, if both cooperate each will get the material reward of $\gamma(1-\sigma)$. The total payoff to each player in each period is the total gain (or losses) in all games he and the other player choose to play. The games differ from each other only by σ, which is uniformly distributed from 0 to 1, and clearly each player prefers to play as many games as possible if the other is expected to cooperate.

A representative game is presented in figure 8.2. To preserve its PD nature, it is assumed that $\beta > 0$ and $\gamma < \alpha$ so that one's best response is D to any action taken by the other. Ignore for the moment the social nature of the game (specifically, the Fs in the payoffs).

Figure 8.2: Social Prisoners' Dilemma Game

	Cooperate	Defect
Cooperate	$\gamma(1-\sigma)+F$, $\gamma(1-\sigma)+F$	$-\beta, \alpha$
Defect	$\alpha, -\beta$	0, 0

This game is infinitely repeated and the players discount the future by the time discount factor of δ. As elaborated upon in chapter 4, some cooperation can be sustained in this game under the threat of terminating future cooperation. Specifically, for a given discount factor, each player will find it best to play C as long as $\sigma \leq (\gamma - \alpha + \alpha\delta)/\gamma$. In all games satisfying this condition, cooperation is self-enforcing when each player believes that the other will cooperate but only as long as none of them defected in the past.[12] Cooperation is possible only in games in which the gains for doing so, relative to the amount that can be gained from cheating, is sufficiently large.

To capture the social aspect of the situation assume that friendship can evolve in the way postulated by Homans and presented above. The repeated cooperation based on economic incentive alone increases the intensity of interactions which leads to feelings of friendship among the interacting individuals. Specifically, cooperation based only on economic incentive increases each player's payoff by the level F, capturing the non-material payoff from having your friend acting in a manner consistent with your expectations of him. Clearly, we can also extend the game to capture disappointment, etc. but this is not necessary to make the point. Similarly, we could have incorporated in the original analysis of the equilibrium that the players' expected friendship to evolve, and that would have influenced their choice of actions in the beginning. Such further modification would have strengthened the result, which is that the player can now cooperate more than before.

Specifically, friendship based on economic incentives enters into the payoffs of the economic game. Cooperation among particular individuals leads to a social - friendship - relationship that supported further economic activity by enabling the players to commit to more than would have been possible otherwise. Once the system ceases to evolve it reaches the

[12] For ease of presentation I ignore the possibility of extending cooperation by conditioning cooperation in one game on others. Such an extension would not have qualitatively changed the results.

equilibrium level of friendship F^* and Cooperation is an equilibrium in any game for which $\sigma \leq (\gamma - \alpha + \alpha\delta)/\gamma + F^*(1 - \delta + \gamma)/\gamma$. More economic cooperation is feasible due to the social relationships among the two players.[13]

8.4 Who Needs You Anyway? Social Ostracism

The psychological stimulus-response mechanism was crucial to the above arguments. This mechanism was central to the credibility of the threat that particular actions would lead to social rewards or sanctions. Social rewards and sanctions do not require action by those who give or inflict them. Indeed, they may only exist in the mind of the recipient. Hence, rewarding is costless while sanctioning cannot be avoided even if it is painful for the one who "inflicts" it due, for example, to the loss of friendship.

Game theory, enables us to explore and deductively restrict arguments regarding another mechanism through which social rewards and sanctions operate. Social rewards and sanctions often transpire in different ways. People actually take actions that express their approval or disapproval of others' actions. In part II we considered actions that influence the material well-being of an individual through their economic and physical implications. The actions of concern here are those that influence one's welfare through their impact on the consumption of "social goods,"such as statues, appreciation, friendship, and the feeling of belonging. The consumption of such goods reflects humans as social animals who gain utility from positive social interactions. Social ostracism thus entails a penalty. It can express itself in such actions as not inviting someone to a party, spiting on the ground in front of someone, or prohibiting one's children from playing with other children. The operation of this mechanism raises the issue of the credibility. Why is the threat of social ostracism credible? Ostracism is credible when a group is sufficiently large so that excluding one member from social interactions does not inflict much if any cost on the other members. Indeed, it may also be a way for them to gain by reinforcing their sense of belonging to the group.

The following is an example of such a mechanism: it is postulated that a group of

---

[13] Because $F(1 - \delta + \gamma)/\gamma$ is positive.

individuals is simultaneously interacting economically and socially.[14] The economic interaction is such that cooperation in, for example, the provision of public good, is not self-enforcing. Social interactions within the group are valuable to each individual but the participation of each particular individual in these interactions is not valuable to other members of the group. Hence, every member can be motivated to cooperate in the economic exchange, expecting that failure to do so will lead to social ostracism, that is, exclusion from the benefits of the social exchange game. This threat, in turn, is credible because this exclusion does not reduce the benefit from the social interaction of other members of the group.[15]

To see exactly how such a mechanism can work, consider a game of public good provision. An important characteristic of such provisions (a park, clean air, etc.) is that no one can be excluded from consuming them. In this game there are $n \geq 2$ players, each of whom has the endowment of y and who decides simultaneously on their contribution levels, q_i, of either zero or q to the public good. The benefit to each player from the public good is P times the total amount contributed minus the amount that he contributed. That is, player i's payoff is

$$U_i(q_1,...,q_n) = y - q_i + P\sum_{j=1}^{n} q_j$$

If P is larger than 1/n, it is optimal that each individual contribute q. If each contributes nothing, each gets the payoff of y. If they all contribute q, each gets y - q + Pnq, which is bigger than y if $P > 1/n$. But if P is less than 1, it is optimal for each player that everyone else contributes while he contributes nothing. This is the case because one's return on his own contribution is $P < 1$. Contributing when everyone else contributes entails y - q + qpn, while not contributing in this case implies y + qP(n - 1) = y - qP + qPn. In this case then, one's best response is to contribute nothing and to hope to free-ride on the contributions of others. But

---

[14] Hirshleifer and Rasmusen 1989 provide an example of another mechanism in which one is punished even if his participation in the social game is beneficial. Each period individuals play the economic and then a social game. The players' strategies call each to not play the social game with anyone who cheated in the previous period or failed to punish someone who was supposed to be punished.

[15] Indeed, exclusion may even have beneficial social side effects to the members of the group, such as providing a topic for conversation and gossip and reinforcing their common identity as those who act "appropriately."

because this holds for everyone, no public good is provided and everyone is worse off.

The situation is changed if we assume that the game is repeated an infinite number of periods and each player discounts the future by the factor of δ. In this case, if the players are sufficiently patient and the total contributed is observable, there is a sub-game perfect equilibrium in which each individual contributes. Each player's strategy is to contribute q as long as long as everyone has always contributed and never to contribute otherwise. The threat to respond to non-contribution by never contributing again is credible because if one expects others not to contribute, one's best response is not to contribute either.

Yet, sustaining cooperation in this manner can be very costly. One bad apple can spoil the whole pie. It is enough for one individual to discount the future less than the others for cooperation to forever cease. Furthermore, when the nature of public good is such that its lack is very costly, as is the case when providing defense, security, flood control, or a water system, the threat of ceasing contributions may not indeed be credible (even if each player's contribution is observable). It would be difficult for members of the group not to argue to let bygones be bygones and renew cooperation.[16]

Social interactions, however, may resolve this problem and enable the group to credibly commit to punish one who did not contribute without resorting to the threat of reducing their own contributions.[17] To see how this can be done, suppose that the same group of individuals is also engaged in a repeated social exchange game similar to the ones described above. Without going into details, the game's important characteristics are that each individual contributes the observable amount of C_s to it and the resulting social private benefits, such as social esteem, approval, and friendship, contribute B_s to that individual's well-being. If one does not contribute in some period, this behavior will be known to everyone in the next period. Furthermore, even if one contributes, the other contributors can exclude him from getting the benefits of the social good.

---

[16] In other words, the equilibrium in not renegotiation-proof. Roughly speaking, if individuals are supposed to punish each other, they face the temptation to negotiate to instead adopt a Pareto-improving strategy. The expectation that this will be the case, undermines the credibility of the punishment to begin with. See discussion in Fudenberg and Tirole, 1991, 174-81.

[17] This specification is similar to the one in Aoki 2001: 47-9.

B_s depends on the number of individuals contributing to the production of the social good and it increases with the number of participating individuals. So $B_s(n)$ is increasing in n. However, suppose that because each individual can socially interact with so many others, $B_s(n)$ reaches its maximum at some $\tilde{n} < N$. Having more than $\tilde{n}$ individuals interacting in the social game does not add to the social benefit that each participants enjoys.

If individuals sufficiently value the future and the benefits from social exchange, each can be motivated to contribute to its production. Specifically, suppose that each member of the group plays the strategy of contributing each period but excludes anyone from the resulting social benefit who did not contribute. If it is believed that such a strategy will be followed, each individual finds it best to contribute if the benefit of getting the social good every period forever, that is, $(B_s(\tilde{n}) - C_s)/(1 - \delta)$, is bigger than the one-time gain from not contributing and getting only the one-time benefit of $B_s(\tilde{n})$. Reorganizing this term implies that one's best response to the above strategy is to contribute if $C_s \leq \delta B(\tilde{n})$.

Now consider the combined influence of the two interactions - the economic and the social. Specifically, suppose that the two games are played simultaneously each period. As we have seen, cooperation in the economic interactions cannot be supported when considered in isolation. Can it be supported when the two interactions are considered? The reason it can be supported is similar to the ones made above: the belief that one will lose future benefits from the social interaction motivates him or her to contribute to the economic one.

Consider the following strategy: Each individual contributes to the public and the social good if and only if he has never failed to do so in the past, and each individual excludes any individual from getting the benefit of the social exchange if that individual ever failed to contribute in either the economic or social interactions. Suppose that the belief that this strategy will be followed prevails. Under what condition will an individual find it optimal to contribute? Note that the threat of punishment is credible because one does not lose anything from excluding another individual from the social exchange game and one who has ever failed to contribute will never contribute again.[18] Given the credibility of the punishment, one's best response is to contribute if the net present value of cooperating every period in both interactions is higher than

[18] Because $N > \tilde{n}$.

the gain from not contributing in a particular period and foregoing these benefits thereafter. As we have seen above, the net present value of cooperation every period in the social interaction is $(B_s(\tilde{n}) - C_s)/(1 - \delta)$. The net present value of contributing to the production of economic, public good is negative, equaling $-(1 - P)q$.[19] The gain from not contributing is $B_s(\tilde{n})$.

Rearranging these terms implies that one would find it optimal to contribute if $\delta B_s(\tilde{n}) - (1 - P)q \geq C_s$. So one would contribute to the economic gain despite the per-period loss of $(1 - P)q$ if there are sufficient offsetting gains from the social exchange, $\delta B_s(\tilde{n}) - C_s \geq (1 - P)q$. Because the two interactions occur at the same time, if there is sufficient enforcement "slack" in the social exchange game, $\delta B_s(\tilde{n}) - C_s$, it can support cooperation in the economic interaction in which the only self-enforcing behavior implies no cooperation.[20]

8.5 Playing Social Games

Game theory provides a framework within which we can examine various inter-relationships between the environment, social relationships, and economic outcomes. This framework has been applied, in particular, to the study of social relationships in organization.[21] But there are also applications for the study of institutions. Clay (1997) has applied this framework to study institutions that govern trade in Mexican California and I present this analysis because it integrates social games and ostracism with economic games and reputation, as discussed in the previous part.

In Mexican California, a group of American long-distance traders was active. They traded a lot among themselves, providing each other with various agency services such as

---

[19] This reflects the fact that when $P < 1$, one does not benefit from his own contribution to the public good.

[20] The term "slack" is from Bernheim and Whinston (1990). They examined the game played by firms in one market and considered the implication on collusion from these firms expanding into another market, thereby gaining the slack required to support more collusion in the first market. By linking two games, more collusion can be sustained. The argument above is similar to such "linked games" analysis, but at its center is the importance of people's inability to determine the dimensionality of their interactions. If individuals could have moved to socially interact with others, there would not have been any slack.

[21] See, for example, Kandel and Lazear 1992, and Gibbons Forthcoming.

handling goods and collecting debts. Honesty in these relationships was maintained through the fear of losing future business with the network members. In addition, these America merchants traded with the local Mexicans and much of this trade was done on credit.

Credit relationships with the local peasant population, however, presented an organizational problem. There were no contract enforcement institutions. Specifically, contract enforcement within the local villages was achieved by social pressure. Disputes were negotiated away. The American merchants, however, were not members of the village community and could not take advantage of these contract enforcement institutions. A solution had been that in villages important to their trade, an American trader would settle down and integrate into the local community. He would marry a local girl, convert to Catholicism, speak Spanish, and raise his children as the locals did. By becoming a member of the community, such a trader had access to the local contract enforcement institution while retaining his affiliation with the American merchants' network.[22]

The example illustrates how the game-theoretical framework contributes to studying the importance and inter-relationships of various institutional elements when motivation is provided for the human propensity to value social relationships. The relatively small Mexican villages constituted groups with an internal enforcement ability based on social relationships. They were taken as exogenous by each of the interacting individuals, merchants and peasants alike. The peasants' immobility implied that they had to take this group and their membership in it as exogenous. At the same time, it was their frequent interactions and presumably information transmission through gossip within the village that generated the intra-village enforcement ability. The village was an organization that altered the rules of the game relevant to each of the interacting individuals while it was endogenous to the actions of all its members.[23]

---

[22] See also the applications and relevant discussion in Bernstein 1992 (regarding contemporary diamond traders); Ostrom, Gardner, and Walker 1996 (regarding communal regulation of common pool resources); Besley and Coate 1995 (regarding the importance of social relationships in motivating the repayment of loans in developing countries); Bernheim 1994 (regarding conformity as reflecting desire for social esteem).

[23] Its essence, however, is distinct from that of the Maghribi traders network in which enforcement was based more on material motivations.

Social immobility, the high cost for individuals to sever their personal relationships with others, is crucial for enforcement based on social relationships. This is well reflected in the actions of the American traders who had to make an irreversible investment in the social and personal relationships in the village to increase the cost of their mobility to the point at which they became part of the local community.

For social relationships to be part of an institution leading to a particular economic behavior, the appropriate cultural beliefs, however, must also prevail, and rules distributing and propagating knowledge regarding the meaning of various actions and the related expected behavior may also be required. The reasons are essentially the same as those discussed in part II. Social relationships, in and of themselves, do not guarantee particular behavior even if it is efficient. In the examples discussed above, the discussion concentrated on a particular equilibrium but others may prevail as well. In the public good game discussed in section 8.4, for example, zero contributions can also be an equilibrium outcome. The game-theoretic analysis, however, enables us to examine what behavior can prevail in a given environment and how, once particular beliefs prevail, they will be regenerated and confirmed by actual behavior.

The game-theoretic analysis of the stimulus-response mechanism has also considered the appropriate behavior, the behavior expected from each individual, to be exogenous to him. Yet, it derives this behavior endogenously as emerging through the interactions among all the interacting individuals. Each individual, taking the actions of the others as given, considered the trade-off between contributing more or using the money for private good consumption. Actions by individuals, each of whom takes the expected actions of others as given, determine the equilibrium amount provided by each.

The game-theoretic analysis of the behavioral implications of social relationships draws attention to, and enables the study of, the role of organizations in influencing these implications. Recall that organizations are man-made, non-technological factors that influence behavior while being exogenous to each of the individuals whose behavior they influence. Organizations that foster social, face-to-face interactions provide the conditions required for social relationships to influence the set of self-enforcing behavioral cultural beliefs. Organizations serving such a function can be informal social structures such as villages, tribes or ethnically-based business groups. But formal organizations, such as firms, churches, military units, Parent-Teacher

Associations, and bowling clubs, can also serve this function.

For example, Ellickson (1991) has presented several case-studies indicating that close-knit groups develop norms of cooperation and dispute resolution that are welfare-maximizing and adhered to by members of the group. Whalers during the nineteenth century, for example, were members of a few intimate and socially interlinked communities. This social familiarity seems to have motivated them to adhere to social norms that regulated their behavior on the high seas in a welfare-maximizing way. Landa (1994, chapter 5) has examined the operation of Chinese middlemen networks engaged in the marketing of smallholders' rubber in Singapore and West Malaysia in the 1960s. She noted that this network "was dominated by a middleman group with a tightly knit kinship structure from the Hokkien-Chinese ethnic group... [among whom] mutual trust and mutual aid formed the basis for particularization of exchange relations" (p. 101).

Empirical evidence also indicates the effect of more formal organizations on the ability to resolve collective actions and advance cooperation, based on internalized institutional elements reflecting social relationships and social preferences. Putnam (1993) has examined the number and nature of voluntary organizations in various parts of Italy. He documented that in areas with more of these organizations, the local government functions better in serving political and economic needs. The ability of members of such organizations to overcome collective action problems enabled them to press the local government to serve them better.

Comment regarding social relationships and economic outcomes:

Another way in which social factors have been integrated into economic analysis is through the way that social statues are applied. (Cole, Mailath, and Postlewaite 1992.) Because the analytical framework is not game-theoretic, this line of analysis falls outside the scope of this work. Nevertheless, a short note is in order because of its social focus and merit. The basic idea is that different societies bestow social status upon their members in different ways. Money, beauty, knowledge, physical strength, and entrepreneurship, for example, can be socially rewarded in different ways. Individuals strive for social esteem and hence the way it is bestowed upon them motivates them to invest their resources in a distinct manner. Cole et al. integrated this idea in a growth model to demonstrate that the different allocation mechanisms of such

statues imply distinct trajectories of economic growth.

&Chapter 9 We Have Feelings Too! Social Preferences

For most of us, a dollar taken from a blind begger does not have the same utility value as a dollar found on the street. Indeed, many of us have willingly given up money earned through labor to increase the welfare of non-kin. At times, we get angry at those who act "inappropriately" and are willing to retaliate even if it does not make economic sense. Arguably, such behavior reflects the human propensity to have feelings toward others and to act emotionally.

Yet, in economics, one's preferences have been traditionally defined over his consumption of material goods and his work effort. The corresponding analysis is thus postulated that people are selfish and materialistic: They are motivated exclusively by considering their own material self-interest. (Henceforth, selfishness.) In such a formulation voluntary acts of altruism, giving to charity, or laboring for the public good is a puzzle (Dawes and Thaler 1988). Indeed, it is even more puzzling to note that people do not harm others despite the ability to gain materially from doing so (Field 2001). Yet, casual observation indicates the importance of such behavior. Parents sacrifice personal wealth and exert great effort in raising their children while we routinely do not harm others to materially benefit ourselves.

Game theory has contributed to our ability to identify the details of such behavior and model it. As a theory of behavior in strategic situations, game theory makes it possible to design experiments revealing how social preferences - people's positive or negative concerns about the material payoffs of relevant others - influence behavior. These experiments, in turn, provide the knowledge required to try to specify a utility function incorporating selfish and social preferences and which can consistently account for the experimental results.

Sections 9.1 and 9.2 present relevant findings from these experiments. Section 9.1 presents the findings that lead to an attempt to capture social preferences as reflecting, in particular, altruism, concern with social welfare, and inequality-aversion.[24] Section 9.2 emphasizes the findings leading to considering one's concern with another person's material

[24] The section presents three utility-function specifications. For others, see Bolton 1991, Bolton and Ockenfels 2000, Kirchsteiger 1994 and Levine 1998. See further discussion in section 9.2.

payoff as reflecting reciprocity and an emotional response to the latter's actions. Game-theoretic experiments also enable considering whether individuals with social preferences act strategically or not. Section 9.3 reviews the evidence which indicates the importance of strategic behavior.

9.1 Social Preferences: Altruism, Inequality, and Social-Welfare

Are people altruistic in the sense that they take pleasure in increasing the material payoffs of others? Experiments conducted by Andreoni and Miller (2002, henceforth AM) substantiated the importance of altruism: It indicates that many people are concerned with increasing social welfare even if it implies a reduction in their own payoffs. In other words, the first partial derivatives of an individual's i utility function, $U_i(x_1, \ldots, x_n)$ are strictly positive. Indeed, some people were willing to let someone else take all the material payoff if it maximized the total surplus. The experiments that AM conducted are a modified version of the Dictator Game (DG).

In the original Dictator Game, developed by Forsythe, et al. (1994) one individual, the dictator, could divide m dollars among himself and someone else. That is, this sum of his payoff, π_s, and that of the other, π_o, has to equal m, $\pi_s + \pi_o = m$. Note that the dictator could have assigned the total amount to himself.

AM experimented with a modified structure of the Dictator Game. In their formulation, the dictator faced different prices for transferring money to the other player. In other words, the dictator had to give up less, the exact amount, or more than one dollar for every dollar that the other player received. The budget constraint that the dictator faced now was $\pi_s + \pi_o = m$. But the income to the other player was $p\pi_o$. Hence, $p > 1$ implies that for every dollar the dictator gave up, the other player got more than a dollar. In this case, providing the other with more implies increasing the total social welfare.

How did the dictators behave? Roughly speaking, 47.2 percent of them acted in a selfish manner, always taking the whole amount for themselves. Denoting a dictator's utility by U_s (where the s is for self) and π_s and π_o the material payoffs for the dictator and the other player respectively, the dictators in this group behaved as if their utility function was of the form $U_s(\pi_s, \pi_o) = \pi_s$. Another 30.4 percent divided the total monetary payoffs equally among the two players, implying a Leontief preference of $U_s(\pi_s, \pi_o) = \text{Min}\{\pi_s, \pi_o\}$. The last 22.4 percent allocated the money in a way that maximized the total monetary rewards, implying that their

preference exhibited a perfect substitute, $(\pi_s, \pi_o) = \pi_s + \pi_o$. Hence, 57 percent of the participants revealed some sort of social preference where some of them seemed to have a notion of Rawlsian (Leontief) fairness, and others seemed to have a Utilitarian (perfect substitute) notion.[25]

While clearly indicating the importance of social preference, there is a debate over whether such results are conclusive evidence for the general importance of altruism (Fehr and Schmidt 2001, section 4.3). Among the reasons is the observation that in other experiments individuals took actions that reduced the welfare of others, did not maximize the total surplus, and responded to the perceived intentions of actions taken by relevant others.

Many experiments suggest that some people care about the equality of the payoffs between themselves and others. They exhibit inequality-aversion. Such aversion is reflected, for example, in behavior in the Ultimatum Game. Like the Dictator Game, in the Ultimatum Game there is a proposer who can suggest how to divide a fixed amount between himself and the responder. The responder, however, can either agree and then the amount is divided according to the proposal, or disagree in which case both get nothing. If the players are motivated only by self-interest, the unique subgame perfect equilibrium is one in which the proposer makes the smallest possible offer which is accepted by the responder.

Numerous experiments were conducted to evaluate this prediction and they were performed in different countries, with different monetary amounts and different experimental procedures. A robust conclusion from these experiments is a rejection of the above prediction. For example, surveying the results of many experiments, Fehr and Schmidt (1999, henceforth FS) have noted that 71% of offers were in the interval of [.4, .5] of the total amount. Note that once again some, but not all, individuals exhibit social preferences.

FS have suggested, therefore, a specification of preferences in which inequality aversion motivates individuals to act in this way. It captures that some people are willing to give up some material payoff to move in the direction of more equitable outcomes. Their inequality aversion is self-centered in the sense that people do not care per se about inequality that exists among

---

[25] AM also argue that a CES utility function provides the best empirical fit for their findings and captures all preferences. $U_s = (\alpha \pi_s^{\rho} + (1 - \alpha)\pi_o^{\rho})^{1/\rho}$ where the parameter α indicates selfishness and ρ captures the convexity of preferences through the elasticity of substitution $\sigma = 1/(\rho - 1)$.

other people but are only interested in the fairness of their own material payoffs relative to the payoffs of others.[26]

Specifically, in FS formulation, one is positively concerned with his own material payoff but negatively concerned with inequalities. This latter concern is asymmetric. One's loss of utility from inequality is higher if it implies a disadvantage to himself than if it implies a disadvantage to the other. That is, one "suffers" more from inequality that is to his disadvantage. Formally, in the case of two individuals, denote by x the vector of material, monetary payoffs to the two players, x_i and x_j. Player i's utility function is $U_i(x) = x_i - \alpha_i \text{Max}\{x_j - x_i, 0\} - \beta_i \text{Max}\{x_i - x_j, 0\}$ where $\alpha_i \geq \beta_i$ and $1 > \beta_i \geq 0$.[27] Note that this specification captures that one is loss-averse in social comparisons: negative deviations from the reference outcome count more than positive deviations.[28]

This simple formulation of inequality aversion accounts for the puzzle of relatively equal outcomes in the Ultimatum Game. By rejecting an unequal proposal, the responder foregoes the utility gains from the monetary reward but accepting an unequal proposal implies a utility reduction due to the implied inequality. Therefore, if offered too little, the responder is better off by rejecting and getting the payoff associated with the equal monetary reward of zero. Anticipating this response, the proposer is better off by making a relatively equal offer to begin with. In other words, if the responder is inequality- averse, proposing an almost equal distribution is an outcome associated with subgame perfect equilibrium.[29] Similar results hold if

---

[26] For similar alternative formulations, see Loewsenstein, et. al. 1989; Bolton and Ockenfels 2000.

[27] For a set of n players indexed by i, the utility function of player i is given by $U_i(x) = x_i - (\alpha_i/(n-1))\sum \text{Max}\{x_j - x_i, 0\} - (\beta_i/(n-1))\sum \text{Max}\{x_i - x_j, 0\}$ where the summation is over all $i \neq j$. The assumption $\beta_i \geq 0$ implies that no one wants to be better off than others but his assumption can be relaxed. β_i is restricted to be less than 1 to capture the idea that one is not willing to throw money away to reduce inequality.

[28] The importance of loss aversion has been stressed by Tversky and Kahneman 1991, among others.

[29] Formally, denote the proposer's preference parameters by $(\alpha 1, \beta 1)$ and those of the responder by $(\alpha 2, \beta 2)$. The following can be established for the case in which the responder's preference parameters are known. In a game in which these parameters are common knowledge, the subgame perfect equilibrium is proposing s* and accepting where s* = .5 if $\beta 1 > .5$, s* is between = .5 and $\alpha 2/(1 + 2\alpha 2)$ if

we introduce uncertainty regarding the responder's preference parameters, (α_2, β_2). In this case, however, rejecting offers would be observed on the equilibrium path. (FS, proposition 1.)

To evaluate the merit of this specification, FS examined whether it can account for behavior in other experiments. It was indeed found to account for behavior in market games in which the outcomes are highly inequitable. An example of such a market game is a situation in which many price-setting sellers (proposers) can sell one unit of a good to a single buyer (responder) who demands only that much. In experiments, the buyer was able to gain all the surplus. This result, however, is consistent with the equality-aversion utility specification. Intuitively, this is the case because in a market setting, unlike the Ultimatum Game, no single player can enforce an equitable outcome. Competition renders fairness consideration irrelevant when the competing players can punish the monopolist by destroying some of the surplus, thereby generating a more equitable outcome.

Charness and Rabin (2001, henceforth CR) have proposed integrating the concern with the social-welfare that AM (and CR) found in their experiments with the inequality-aversion model of FS. They were particularly motivated by their observation that participants in their experiments were willing to give up some of their material payoff to increase the payoffs for all recipients, especially low-payoff recipients. Individuals make *inequality-increasing* sacrifices when these sacrifices are efficient and inexpensive.[30]

CR proposed the following utility function formulation that captures inequality aversion and concern with social welfare. Let π_A and π_B be the two monetary payoffs and let $U_B(\pi_A, \pi_B)$ denote B's utility. Specifically, $U_B(\pi_A, \pi_B) = \rho\pi_A + (1 - \rho)\pi_B$ if $\pi_B > \pi_A$ and $U_B(\pi_A, \pi_B) = \sigma\pi_A + (1 - \sigma)\pi_B$ if $\pi_B < \pi_A$.[31] The parameters ρ and σ capture various possible social preferences. The selfish case is captured when $\sigma = \rho = 0$, implying that $U_B(\pi_A, \pi_B) = \pi_B$, while the case in which B wants to do as well as possible in comparison to A is captured when σ and ρ are both negative and $\rho \geq \sigma$. The FS inequality- aversion specification is captured when $\sigma < 0 <$

$\beta 1 = .5$, and $s^* = \alpha 2/(1 + 2\alpha 2)$ if $\beta 1 < .5$. See proposition 1 in FS.

[30] Similar results were found by Charness and Grosskopf 2001, and Kritkos and Bolle 1999.

[31] The above can be expressed as $U_B(\pi_A,\pi_B) \equiv (\rho r + \sigma s)\pi_A + (1 - \rho r - \sigma s)\pi_B$ where r = 1 if $\pi B > \pi A$, and r = 0 otherwise; s = 1 if $\pi_B < \pi_A$, and s = 0 otherwise.

$\rho < 1$. That is, B likes a high monetary payoff and prefers that payoffs are equal, including the wish to lower A's payoff when A does better than B. Social welfare preference is captured when $1 \geq \rho \geq \sigma > 0$. Here, one always prefers more for himself and the other person, but is more in favor in getting payoffs for himself when he is behind than when he is ahead.

9.2 Social Preferences: Reciprocity and Caring About Others' Intentions

The previous section presented the attempt to understand social preferences - people's positive or negative concerns about the material payoffs of relevant others - as reflecting one's **unconditional** concern with the welfare of others. This concern was unconditional in the sense that the way in which one was postulated to care about the welfare of others did not depend on their past actions or his perception of their intentions.

Experiments in various games, however, have indicated that social preferences are responsive to past actions and the perceived intentions of the others. Individuals acted to raise or lower others' payoffs, depending on the actions that these others took in the past and what the judgment was regarding their intentions.

In particular, experiments indicate the importance of **reciprocity**. Many responded to behavior deemed to be "fair" by similar actions that raised the other's material payoffs. Indeed, they were willing to forego material reward to increase the welfare of those who acted fairly toward them. At the same time, many people are revengeful, willing to reduce their own material payoffs to reduce the material welfare of those who have acted unfairly toward them.

People's willingness to punish others for what they consider to be unfair behavior is well reflected in comparing results in the Dictator and Ultimatum Games. Recall that in both games the proposer can suggest an allocation of a fixed sum between himself and another person. In an Ultimatum Game, however, the responder has the option of rejecting the offer. If people are motivated only by altruism or inequality-aversion, the outcome in both games should be the same. If people are reciprocators, this should not be the case. In particular, if people are willing to punish others for what they consider to be an unfair - very low - offer and this is anticipated by the proposers, we would observe that higher offers were being made in the Ultimatum relative to the Dictator Game.

Forsythe et. al. (1994) compared the two games and found that indeed, offers were

substantially and significantly higher in the Ultimatum Game. In a $10 Dictator Game, 21 percent of the proposers gave the other nothing and 21 percent gave the other an equal share. In a $10 Ultimatum Game, however, all proposers offered the other a positive amount and 75 percent offered at least an equal amount.

Experiments in public good games provide a much larger body of evidence that people reciprocate. Fischbacher, Gächter and Fehr (2001) conducted experiments in public good games and found that 50 percent of the participants were conditional cooperators. Their contributions positively increased with the average contributions. The willingness to forego material reward to retaliate against unfair behavior toward them was also studied in the context of public good games.

Consider a regular public good game identical to the one discussed in section 9.1. There are $n \geq 2$ players, each of whom has the endowment of y and who decided simultaneously on his contribution levels $q_i \in [0, q]$, $i \in \{1,...,n\}$, to the public good.

$$x_i(q_1,...,q_n) = y - q_i + P\sum_{j=1}^{n} q_j, \quad 1/n < P < 1,$$

where P denotes the constant marginal return to the public good, $Q \equiv \sum_{j=1}^{n} q_j$. As we have seen in 9.1, since $P < 1$, one's marginal investment to public good yields that person a monetary loss of (1 - P). Thus, if social preferences do not matter, it is optimal to each individual to contribute $q_i = 0$. However, since $P > 1/n$, the aggregate monetary payoff is maximized if each player chooses $q_i = q$.

Fehr and Schmidt (1999) summarize the results of many experiments with such public good games. In many of these experiments the game was repeated several times so Fehr and Schmidt examined only behavior in the last period. These experiments were conducted in six countries, involved more than a thousand individuals, the group size ranged from four to sixteen, and P, the marginal pecuniary return ranged from .2 to .75. On average 73 percent of all participants free-rode and contributed nothing. In other words, about a quarter of the players were willing to contribute to the public good in the last period.[32] The range of free-riding in various experiments was from 54 to 89 percent.

---

[32] In repeated settings, contributions in early stages were often higher. Experiments suggest that the decline over time was because those who contributed early retaliated against free-riding by ceasing to contribute. See discussion below.

To identify reciprocity in this context, consider an augmented, two-stage public good game. The first stage is identical to the one described above. At stage 2 each player is informed about the contribution of all other players, namely, the contribution vector $(q_1,...,q_n)$. In the second stage of the game, each player can, simultaneously with others, impose a punishment on any other player. In other words, player i chooses a punishment vector $p_i = (p_{i1},...,p_{in})$, where $p_{ij} \geq 0$ denotes the punishment player i imposes on player j. The cost of this punishment to player i is given by $c\sum_{j=1}^{n} p_{ij}$ where $0 < c < 1$. Player i, however, may also be punished by the other players, which generates an income loss to i of $\sum_{j=1}^{n} p_{ij}$. Thus, the monetary payoff of player i is given by $x_i(q_1,...,q_n,p_1,...,p_n) = y - q_i + P\sum_{j=1}^{n} q_j - \sum_{j-1}^{n} p_{ij} - c\sum_{j=1}^{n} p_{ij}$.

Note that if people are motivated only by social welfare or inequality considerations, the result in this game should be the same as before. Fehr and Gächter (2000a) experimentally evaluated this prediction. Furthermore, the experiment design was also aimed at eliminating cooperation based on the effect of reputation in repeated interactions as discussed in part II.[33]

The results are unambiguous. In the regular (one-stage) public good game, contributions were relatively low. When individuals interacted repeatedly with each other for a finite number of periods, the average contribution in all periods was about 37 percent of the endowment and contributions over time declined to reach 16 percent in the last period. Even less cooperation was achieved when individuals did not interact repeatedly. The average contribution in all periods without punishment was 18.5 percent of the endowment and it gradually declined to reach about 10 percent in the last period.

The results were significantly different, however, when the game was expanded to include stage two. When punishment was possible, the average contribution in all periods when the same individuals interacted repeatedly was 85 percent of the endowment and reached 91 percent at the last period. Even when individuals did not interact repeatedly, the average contribution over all periods reached 57.5 percent of the endowment and stayed at about this level at the last period as well. Punishment of those who contributed, although feasible, was not carried out.

These results support the assertion that there are subjects who are willing to punish free-

[33] Their experiments had a more elaborate punishment structure that is omitted here for simplicity.

riding and their existence is anticipated by at least some potential free-riders. The anticipation that free-riding will be punished prevents it from the beginning. Furthermore, considering individuals' behavior indicates that those who deviated more from average contributions were punished more severely and they responded to this punishment by increasing their contributions. Finally, they were individuals who inflicted punishment to generate an increase in average contributions and they were successful in achieving this.

Other experiments revealed a brighter side of reciprocity: Many people are not only ready to punish those who do not cooperate, but many are also willing to reciprocate for behavior that seems fair or reflects good intentions. In other words, people are willing to forego material rewards to increase the payoffs of those who treat them well. This positive reciprocity is well reflected in Gift Exchange Games (GEG). In these games the proposer offers a wage, w, to the responder. The responder can accept or reject the offer. In the case of rejection, both players receive the payoff of zero while in the case of acceptance, the responder has to make a costly "effort"choice , e. The monetary payoff for the proposer is $x^p = ve - w$, while the responder's payoff is $x^r = w - c(e)$, where v denotes the marginal value of effort to the proposer and c(e) the strictly increasing effort cost schedule. Clearly, if the responder maximizes only monetary payoffs, his best response is always to accept any offer and to choose the lowest possible effort level. The subgame perfect equilibrium thus implies that the wage, w, will be the lowest possible.

Experiments, however, revealed that in general there is a strong positive correlation between the mean effort and the offered wage. This finding is consistent with the interpretation that the responders, on average, reciprocate generous wage offers with a generous effort level. The level of individuals exhibiting such positive reciprocity is frequently about 40 percent of the responders.[34] Experiments thus strongly suggest that some people exhibit reciprocity. They are conditionally cooperative and are willing to engage in the costly punishment of free-riders. Some people are sensitive to the intentions they perceive are behind the actions that other people take relevant to their payoffs.

---

[34] See discussion in Ostrom 1998, Fehr and Fischbacher 2001b, and Charness and Rabin 2001, and the reference provided there which also refers to evidence for reciprocity drawn from the study of other games.

Given the evidence regarding the importance of reciprocity, explicit specifications of social preferences aimed at capturing this effect have been proposed in the literature. For example, the specification of the inequality-aversion model of Fehr and Schmidt (1999) was $U_i(x) = x_i - \alpha_i Max\ \{x_j - x_i, 0\} - \beta_i Max\ \{x_i - x_j, 0\}$ where $\alpha_i \geq \beta_i$ and $1 > \beta_i \geq 0$. As they noted, "the lack of explicit modeling of intentions... does ... not imply that the model is incompatible with intentions-based interpretations of reciprocal behavior. In our model reciprocal behavior is driven by the preference parameters α_i and β_i. The model is silent as to why α_i and β_i are positive. Whether these parameters are positive because individuals care directly for inequality or whether they infer intentions from actions that cause unequal is not modeled. Yet, this means that positive α_i's and β_i's can be interpreted as a direct concern of equality as well as s are reduced-form concern for intentions. An intention-based interpretation of our preference parameters is possible because bad or good intentions behind an action are, in general, inferred from the equity implications of the action" (p. 853).

Indeed, Fehr and Schmidt (1999: 854) summarized experimental results from two types of Ultimatum Games. In the first, the amount proposed is determined randomly while in the second case the proposer determines how much to offer. The rejection threshold in the second type of game was much higher than in the first type, suggesting that indeed, the parameter α_i shifts as a result of the structure of the game. If one is not responsible for the amount offered, one has no bad intentions in shifting the α_i downward, leading to a lower acceptance threshold.

Similarly, the full specification of the utility function suggested by Charness and Rabin (2000) includes a shift parameter θ which reflects reciprocity. Specifically, they assume that $U_B(\pi_A,\pi_B) \equiv (\rho r + \sigma s + \theta q)\pi_A + (1 - \rho r - \sigma s - \theta q)\pi_B$, where r = 1 if $\pi_B > \pi_A$, and r = 0 otherwise; s = 1 if $\pi_B < \pi_A$, and s = 0 otherwise; q = -1 if A has misbehaved, and q = 0 otherwise.[35]

Game theory enables going beyond considering people's feeling or emotional responses to others as a shift parameter, as will be presented in the next chapter.

---

[35] Alternatively, $U_B(\pi_A, \pi_B) = (\rho + \theta q)\pi_A + (1 - \rho - \theta q)\pi_B$ if $\pi_B \geq \pi_A$ and $U_B(\pi_A, \pi_B) = (\sigma + \theta q)\pi_A + (1 - \sigma - \theta q)\pi_B$ if $\pi_B \leq \pi_A$ where q = -1 if A has misbehaved and q = 0 otherwise.

9.3 Rational Behavior and Social Preferences

The above experimental evidence unambiguously reveals the partiality of the neo-classical specification of utilities. Individuals often do not behave in a manner implied by the assumption that they are motivated only by seeking to improve their own material well-being. Is it then still appropriate to consider individuals as rational? Clearly, it depends by what one means by the term rationality. The appropriate meaning to consider at this point, given the focus of this and the previous part, touches upon decision-making in strategic, economic situations. Do individuals have stable preferences regarding outcomes in such situations?[36] Are they motivated by the consequences of their actions? Do they act strategically?

There are three ways to use experimental results to address these questions: First, to consider if there is a non-rational explanation that better fits the data; second, to test if behavior is consistent with some well-behaved preference ordering; third, to use experimental results to evaluate whether people are motivated by consequences and behave strategically. The evidence available so far is inconsistent with non-rational accounts, consistent with a well-behaved preference ordering, and reflects consequential and strategic behavior.[37]

A non-rational explanation accounting for experimental evidence has been that it reflects selfish motivation and learning.[38] For example, Roth and Erev (1995) and Binmore, Gale and Samuelson (1995) try to explain the presence of fair offers and rejection of low offers in the

---

[36] For survey of psychological evidence indicating that individual's do not always have stable preferences, see Rabin 1998. There are two main issues. First, people have difficulties evaluating their own preferences: they don't always accurately predict their own future preferences or even accurately assess the well-being they have experienced from past choices. Second, research on framing effects, preference reversals, and related phenomena revealed that people may prefer some option x to y when the choice is elicited one way, but prefer y to x when the choice is elicited another way. The first issue is more relevant to what people want to exchange and less relevant to the issue here, namely, institutions enabling exchange. The second issue is part of the analysis discussed in chapter 10. Institutions, for example, frame the context of the exchange.

[37] E.g., Hoffman et al. 1994, Fehr and Schmidt 2001a, Henrich et al. 2001a, Falk and Fischbacher 2000.

[38] Another explanation is that behavior in the laboratory reflects the rule-of-thumb. Because most our real life interactions are repeated, we act in the laboratory as if it is still the case (Hoffman et al. 1996). This may be partially true but cannot account for many of the results reported above such as the tendency to cooperate when interactions are anonymous and behavior is known to be of short duration.

Ultimatum Game by learning models that are based on purely monetary considerations. The central idea is the distinct incentive to learn for the Responders, who can either accept or reject an offer, and the Proposers, who determine how much to propose. For responders, a rejection of low offers is not costly and hence they only learn slowly not to reject them. But such rejections are very costly to the proposers who therefore quickly learn not to make them. Hence, behavior may not converge to the subgame perfect equilibrium in which the lowest possible offers are made. The validity of such learning arguments with respect to simple games such as the Ultimatum Game, however, seems doubtful. Furthermore, in many studies as further discussed below, proposers do anticipate the reaction of the responders appropriately.[39]

Andreoni and Miller (2002) constructed their experiments to test whether behavior exhibiting social preference is consistent with well-behaved preference ordering. As described above, their Dictator Game experiments were such that they could change the "price" to the dictator for acting in a manner benefitting the other. In other words, they changed the budget constraint that the dictators faced and hence could examine the behavior of the same individual under different budget constraints. Hence, they could have tested whether individuals' behavior satisfied the necessary and sufficient conditions required for the existence of well-behaved preferences.[40]

Their results were unambiguous. They concluded that preferences are predictable and well- behaved on the aggregate level and individuals exhibited a significant degree of rationally altruistic behavior. Indeed, over 98 percent of the subjects made choices that are consistent with utility maximization. They found that it is indeed possible to capture altruistic choices with

---

[39] The merit of an alternative theory, that individuals act in a one-shot game as they do in repeated games, is discussed below.

[40] Specifically, they have examined whether individuals have preference ordering that satisfies the Generalized Axiom of Revealed Preference (GARP). A is directly revealed as preferred to B if B was in the choice set when A was chosen. If A is directly revealed as preferred to B, B is directly revealed as preferred to C, ... to Y, and Y is directly revealed as preferred to Z, then A is indirectly revealed as preferred to Z. The GARP is: If A is indirectly revealed as preferred to B, then A is not strictly within the budget set when B is chosen, that is, B is not strictly directly revealed as preferred to A. Satisfying GARP is both a necessary and sufficient condition for the existence of well-behaved preferences, given linear budget constraints.

quasi-concave utility functions for individuals - altruism is rational.[41] Furthermore, Andreoni and Miller found that a model capturing the preference revealed in one experiment consistently accounts for behavior in other experiments. Similar results are reported in Fehr and Schmidt (1999).

Many experiments revealed that individuals respond as postulated in game theory to the strategic environment within which they interact (e.g., Forsythe, et. al. 1994).[42] In hundreds of double- auction experiments, prices and quantities quickly converged to the competitive equilibrium predicted by standard self-interest theory.[43] Backward induction is well reflected in Ultimatum Games as many proposers seem to anticipate that low offers will be rejected with a high probability. Recall, for example, the comparison of the results of the Dictator Games (DG) and Ultimatum Games. In a DG the responder's option to reject is removed - the responder must accept any proposal. Forsythe et. al. 1994 was the first who compared the offers in UGs and DGs. They report that offers are substantially higher in the UG, which suggests that many proposers do apply backwards induction.

Similar results are reported in the cross-country analysis in Roth et al. (1991) and Henrich et al. (2001a). In the latter study, experiments were conducted in fifteen very different societies and the researchers concluded that in all of them individuals exhibited stable preferences and behavior motivated by consequences. Indeed, in each society people, by and

---

[41] As they note, however, their analysis did not explore the influence of the changing environment - the rules of the game, level of anonymity, the gender or age of the participants, or the framing of the decision - on the preference ordering.

[42] Ostrom 1998, however, argues that "what is clearly the case from experimental evidence is that players do not use backward induction in their decision-making plans in an experimental laboratory" (p. 5). The context of these words, however, suggests that what she might have had in mind is that the results are inconsistent with backward induction in finitely-repeated games under the assumption that people are motivated only by self-interest. In any case, she refers to two papers to support the above position, Rapoport 1997, and McKevey and Palfery 1992. Rapoport's analysis, however, was not concerned with rejecting backward induction and his focus and main conclusion were regarding the importance of the framing effect on behavior. The framing effect was captured by information about the order of play (p. 133). He notes that order of moves influences equilibrium selection. McKevey and Palfery 1992 examined the centipede game, which is problematic as far as backward induction is concerned. (Fudenberg and Tirole 1991: 96-100.) In any case, they concluded that a game of incomplete information based on reputation explains their data.

[43] See surveys in Davis and Holt 1993 and Hagel and Roth 1995.

large, correctly anticipated the responses of others. Backward induction was also found by Fehr and Schmidt (1999) who reported that in twelve public good games without punishment where free-riding is a dominant strategy, average and median contributions in the first period were between 40 to 60 percent of the endowment, but fully 73 percent of the participants contributed nothing in the last period.

Individuals seem also to be able to recognize and respond to the strategic difference between one-shot and repeated games. Fehr and Fischbacher (2001a) explicitly tests if individuals understand this difference and the evidence indicates that, by and large, they understand it very well. Fehr and Fischbacher ran two sets of Ultimatum Game experiments under different conditions. In both cases, subjects played against a different opponent in each of the ten iterations of the game. Under one condition, however, the proposers knew nothing about the past behavior of their current responders. Under the other "reputation" condition, the past behavior of the responders was known. If individuals understand the distinction between one-shot and repeated interactions, responders are motivated to build up reputations for "toughness" and rejection of low offers. Hence, the acceptance threshold, the offer that the responder accepts, should increase. Indeed, slightly more than 80 percent of the responders increased their acceptance thresholds in the reputation condition.[44]

Experimental evidence thus lends support to the claim that individuals are rational in the sense of having stable preferences and are motivated by the consequence of their actions. Furthermore, it indicates that people are acting strategically, trying to anticipate others' responses to their actions, adjusting their responsive actions to others' actions, and using backward induction.[45]

It is important to emphasize that the above notion of rationality is very specific. It is concerned with stability of preferences and actions consistent with consequential considerations. Hence, the above discussion does not imply, for example, that individuals are "rational" in the

---

[44] For similar results in Gift Exchange Games, see Gächter and Falk 2002. These findings undermine the suggestion that individuals exhibit dispositional social preferences because they mistake the finite laboratory experiments with repeated, real life situations.

[45] Lindbeck 1997 elaborates on why it is appropriate to consider that individuals act rationally given the values that they have internalized.

sense of having a perfect knowledge of the world around them or unlimited computational capacity. Indeed, below I will argue that individuals do have bounded rationality in this sense. But the above experiments do not shed any light on such rationality considerations because the subjects in them had to choose an action in relatively simple situations and after receiving a detailed description of them. Furthermore, these experiments required identifying and formulating responses to actions taken by other individuals. As the social psychologists (Tooby and Cosmides 1992), have demonstrated, evolution has fine-tuned our brain's capacity to take actions in exactly these situations. Psychological considerations known to cause bias in decision-making, such as inferring too much from too little evidence or conformity bias (e.g., Rabin 1998), are not captured in these experiments.

9.4 Looking Ahead

If some individuals have social preferences, what does this imply regarding institutional analysis?

&Chapter 10 Social Preferences, Norms, Emotions, and Internalized Institutional Elements

Experiments based on game-theoretic analysis have advanced our knowledge regarding how humans' social propensities express themselves in social preferences. Although some individuals behaved purely selfishly in the experiments, others were concerned with the welfare of others and often exhibited concern about social welfare and inequality-aversion. There are also individuals who exhibit reciprocity, who respond to others' behavior by rewarding behavior they consider fair, and are willing to reduce their own material payoffs to retaliate against behavior they consider unfair or mean.

This chapter integrates the discussion of social preferences with that of institutional analysis and the contributions made by the analytical framework provided by game theory. Section 10.1 elaborates on the use of game theory to deductively restrict arguments regarding feasible behavioral cultural beliefs when social preferences are considered, that is, when we consider social preferences as exogenous to the analysis. Section 10.2 considers first the implications on behavioral cultural beliefs reflecting the asymmetric information regarding peoples' social preferences. It then considers the various influences that organizations can have on such beliefs in the presence of asymmetric information.

The discussion in these two sections assumes that social preferences are attributes of individuals that do not depend on the situation. Section 10.3 presents evidence on the limitations of this assumption. Among those who have social preferences, their manifestations are situation-specific. These manifestations systematically differ among societies and within them based on the situation. Presenting this argument in another way, recall that the previous chapter presented social preferences as functions of various parameters. Among these parameters are the relevant others, whose welfare enters into one's preferences, the weight placed on social welfare of certain others relative to one own material payoff, the income that one would allocate among others, and the behavior that triggers reciprocity by being considered fair or unfair. Evidence indicates systematic differences in these parameters among societies and situations.

Section 10.4 therefore argues that the situational contingency of such parameters indicates that they are socially determined. They reflect a process through which society shapes one's utility function. Beyond ones' direct kin, it is socially determined who the relevant others

are whose welfare influences one's utility, the social welfare that enters one's utility, the utility implications of whether or not to divide a particular income among certain others, and what behavior leads to a positive or negative emotional response. Such manifestations of social preferences that are socially determined, that is, they are exogenous to an individual and prevail among members of a particular society, are institutional elements. They can be referred to as internalized (social) institutional elements because they are institutional elements that have been incorporated into one's utility function and reflect social propensities.

Two subsections present the game-theoretic contributions to the study of internalized institutional elements. Subsection 10.4.1 associates the discussion of social preferences to the study of norms. It argues that the parameters in one's social preferences are internalized norms. That is, they reflect the incorporation of a behavioral standard in one's super-ego. The discussion then presents how evolutionary and classical game theory contributes to the study of norms. Game theory restricts assertions regarding norms that can prevail in a given environment and makes explicit the forces leading to the propagation of an internalized norm. Subsection 10.4.2 presents a similar discussion regarding the contributions of psychological game theory to the study of reciprocity. It links it to the study of emotion: the socially determined emotional response to actions by others.

Section 10.5 presents the role of rules and communication in influencing behavior in the presences of social preferences. Section 10.6 concludes by presenting some game-theoretic analyses regarding the origin of social preferences.

It should be noted that similar to part II, the discussion does not present the origin of internalized institutional elements or their implications (such as efficiency or distribution). These issues are examined in subsequent parts.

10.1 Social Preferences, Rules, and Behavioral Cultural Beliefs: A Dispositional Perspective

It is straightforward to extend the analysis of institutional elements presented in parts I and II while taking as exogenous that some people have particular social preferences. Institutional elements are the man-made, non-technological factors that generate regularities of behavior and expected behavior while being exogenous to each of the individuals whose

behavior they influence. To study such elements in the presence of social preferences, we can conduct the analysis presented in parts I and II while replacing the assumption that individuals are selfish with the assumption that some individuals have particular social preferences. Indeed, sections 9.1 and 9.2 presented various specifications of utility functions that represent social preferences in the form of altruism, inequality-aversion, and reciprocity.

Using such utility specifications it is possible to revisit, without using a game, the study of institutional elements, such as the exogenously enforced rules discussed in chapter 2. For example, social preferences influence the economic implications of rules governing property rights allocations. To illustrate this possibility, consider the case of financing a public welfare program using a lottery. The issue is allocating the property rights for the lottery to maximize revenue. Assume that individuals are selfish and the government is less efficient than a firm in managing the lottery. In this case, ignoring the issue of optimal design, it is best to auction the property rights over the lottery and give property rights over the lottery to the winning firm. This maximizes the revenue and minimizes the cost of the lottery.

But this property rights allocation would not be optimal if individuals are altruistic and care about the welfare of those who would benefit from the lottery proceeds. If the government has the property rights, individuals are motivated to purchase lottery tickets, both by the desire to gamble and to benefit the poor. But if a firm has the property rights, individuals are motivated to purchase tickets only by the desire to gamble because the government revenue, and hence the benefit to the poor, has already been determined through the lottery and no longer depends on their purchases.[46] The total revenue from the lottery is therefore lower.

Optimal contractual and organizational design can be similarly revisited while considering the influence of social preferences. Consider, for example, the optimal design of incentives within a firm. If individuals are reciprocators, incentive contracts crowd out voluntary contributions to work. (Frey 1997.) Indeed, Fehr and Gächter (2000b) have presented experimental evidence that such crowding out is observed in laboratory settings. Incentive contracts are chosen not because they are efficient, but because they are more profitable to the

[46] This argument is inspired by a paper by John Morgan, but the only current version of it (Morgan 2000) does not include it.

one who initiates them.[47] Optimal incentive contracts may thus involve gift-exchange: paying workers more than is optimal given their best response in the absence of reciprocity to invoke a reciprocal behavior and a high work effort.

Similarly, we can use game theory to study self-enforcing behavioral cultural beliefs while assuming that individuals have social preferences. Just as in the case of selfish preferences, game theory restricts admissible arguments by exposing the set of self-enforcing cultural beliefs in a given environment. In general, the set of admissible - self-enforcing - behavioral cultural beliefs will be larger. For example, one can be expected to repay a loan if he believes that the lender is a reciprocator who will respond to default by behaving in a violent manner despite the material cost of doing so. If the expected utility loss from such violent behavior is sufficiently high, the borrower will be deterred from defaulting. If lenders in general are believed to be reciprocators, the behavioral cultural belief that loans will be repaid can thus prevail - be self-enforcing - in situations in which they otherwise could not. Social preferences change the set of self-enforcing cultural beliefs.[48]

In exploring the institutional implications of social preferences one has to take into account that, as reflected in experiments, not everyone has social preferences. Hence, studying behavioral cultural beliefs while taking social preferences into account requires also taking into account the heterogeneity in social preferences and the role of organizations in their presence. These issues are examined in the next subsections.

10.2 Incomplete Information and Organizations

An important insight of game theory is that even if the number of individuals with social preferences is small, the knowledge that some individuals have them can greatly impact possible self-enforcing behavior and hence cultural beliefs. Specifically, game-theory situations in which some individuals have an unobservable attribute are examined in game theory using incomplete information models. When applied to the issue at hand, such models assume that it is common knowledge that individuals are of distinct types: some (whose identities are unknown to others)

---

[47] For further applications, see Fehr and Schmidt 2001.

[48] For empirical analyses in this spirit, see Levi 1998 and Bowles et. al. 2000.

have social preferences while some do not. Such incomplete information can have a very large influence on the set of self-enforcing cultural beliefs because individuals may have an incentive to pretend, to act as if they have internalized particular norms or beliefs.

If, for example, acquiring the reputation of being revengeful is sufficiently profitable, one would find it optimal to act as if he is a revengeful type despite the short-term cost implied by someone taking revenge while not being revengeful. By acting like a revengeful type, one who is not that way has to bear the cost of doing so to induce others to update their beliefs about what sort of person he is. What he gains is acquiring the reputation of being revengeful. Hence, although everyone knows that the number of revengeful individuals is actually small, the inability to know who is actually revengeful can lead to everyone being considered as if they are revengeful. The behavioral cultural beliefs that are self-enforcing in this case are those associated with everyone being revengeful.

To give another example, one can find it optimal to cooperate despite the ability to cheat, just to cause others to believe that he has internalized the norm of honesty. In repeated but finite-horizon Prisoners' Dilemma Games, for example, cooperation can be sustained for some periods because individuals find it profitable to gain a reputation of having internalized the norm of honesty. As the end of the game draws near, however, the gains from cheating become bigger than the gains from maintaining an honest reputation and cooperation ceases.

Game-theoretic incomplete information models enable examining the exact conditions under which the above analysis is valid.[49] In general, individuals have more incentives to mis-represent their type for more periods despite the cost involved when the number of periods in which they will interact is larger, the higher is the per-period gain from being considered the other type, and the smaller is the per-period cost of imitating the other type.

Behavior consistent with this insight of incomplete information models was found to prevail in experimental studies. Gächter and Falk (2002) have examined behavior in both one-shot and repeated gift-exchange games. Recall that in these games the proposer offers a wage, w, to the responder. The responder can accept or reject the offer. In the case of rejection, both

[49] The classical reference is to Milgrom, Kreps, Roberts, and Wilson 1982. See also Fudenberg and Tirole 1993.

players receive the payoff of zero while in the case of acceptance the responder has to make a costly "effort" choice, e. The monetary payoff for the proposer is xp = ve - w, while the responder's payoff is xr = w - c(e) where v denotes the marginal value of effort to the proposer and c(e) the strictly-increasing, effort-cost schedule. Clearly, if the responder maximizes only monetary payoff, his best response is always to accept any offer and to choose the lowest possible effort level.

Gächter and Falk examined behavior in two treatments of such a game. In the first, which they refer to as the OS-treatment, the parties are informed that they will never play against each other again. In the second, which they refer to as the RG-treatment, the parties know that they will play ten times. Reciprocity, or a significant, positive, wage-effort relationship, was found in both treatments. Reciprocity and incentives provided by repeated interactions seem to complement each other. The positive wage-effort relationship was steeper in the RG-treatment than in the OS-treatment, and in the RG-treatment, effort levels were higher than in the OS-treatment. Finally, about 50 percent of the individuals who revealed themselves as selfish in the last period by providing the selfish amount of labor imitated the reciprocators in all other periods of the RG-treatment.[50]

Although the logic beyond models of incomplete information is intuitively appealing and the above experimental results indicate their relevance, such models were not widely applied for empirical positive studies. The deficiency of them is their flexibility. By arbitrarily specifying the nature and magnitude of asymmetric information regarding the players' types, any behavior can be generated as a self-enforcing outcome. (Hart 2001.) This deficiency notwithstanding, the experimental evidence indicates the promise of exploring the implications of incomplete information regarding social preferences, such as positive and negative reciprocity and internalized norms.

The discussion so far has ignored organizations, that is, man-made, non-technological factors that change the relevant rules of the game to the interacting individuals while being exogenous to each of them. As we have seen in part II, game theory enables examining the self-enforcing behavior that can prevail in the absence or presence of a particular organization while

[50] For similar findings, see Fischbacher, Gächter and Fehr 2001.

taking into account social preferences. The very same analysis can also be conducted while replacing the assumption of selfish preferences with social preferences. Accordingly, what follows is a short discussion of the various insights regarding the role of organizations in supporting self-enforcing behavior when human social propensities are taken into account.[51]

When there is asymmetric information about who has particular social preferences or has internalized particular norms, organizations can alter this asymmetry by storing and providing the information. They have a similar and yet distinct role from that examined in part II. In part II, information was important in providing a link between past conduct and future reward, which motivated individuals to act in a particular way (e.g., cooperation in the exchange game). When incomplete information about social preferences prevails, organizations serve a similar role. As we have seen, information transmission is crucial for inducing an imitation of a "good type." Organizations that transmit such information thereby change the set of possible self-enforcing cultural beliefs.

But organizations can have an additional role. In situations in which only, for example, reciprocators will cooperate, the information provided by organizations enables the **initiation** of cooperation. Organizations that transmit information about past conduct enable the identification of reciprocators. Cooperation, however, is self-enforcing based on these social preferences. Organizations that provide such information can be informal, such as social networks, or formal, such as employment agencies.[52]

Production technology or the need to cooperate in the provision of public good often requires cooperation among many individuals. Such cooperation, as we have seen, can be based on social preferences. Altruism, inequality-aversion and reciprocity can motivate individuals to cooperate. At the same time, reciprocity can undermine such cooperation. To see why this is the case recall that reciprocity is conditional on the behavior of others. One is willing to reciprocate if others do, but responds to those who do not cooperate by stopping cooperation as well. Furthermore, reciprocators are willing to bear a cost for punishing those who failed to cooperate

---

[51] A further role of organizations in socialization is discussed in section 10.5.

[52] Granovetter 1974 has argued that weak social ties in the labor force are important in attaining a job. Weak social ties link primary groups and hence they exist among individuals who are not competitors in the labor market.

when they did. Hence, the presence of some non-reciprocators can cause reciprocity-based cooperation to unravel.

Such unraveling is well reflected, for example, in experiments in public good games. In repeated experiments, individuals initially contributed relatively a lot but their contribution declined over time in response to the observed failure of others to contribute. Fehr and Gächter (2000a), for example, found that individuals contributed on average 40 percent of their endowment in the first few periods. After ten periods, contributions declined to less than 20 percent. People do not like to let others free-ride on their contributions. Indeed, experiments have confirmed that individuals are willing to spend resources to punish those who fail to contribute.

Organizations can thus play a role in preventing such **unraveling** of **reciprocity**-based cooperation by punishing only those who fail to cooperate. Indeed, Ostrom, Gardner, and Walker's (1992) experimental results suggest this role of organizations. In a repeated public good game, individuals were informed about each others' past contributions and they were allowed to communicate about and contribute to the punishment of others. Punishment was costly to whoever contributed to it. Contributions were 93 percent of the optimal amount and only a few participants defected and were punished. The net benefits after punishing were at a 90 percent level of the social optimum. Similarly, when Fehr and Gächter (2000a) enabled players to spend resources on punishing those who did not contribute, average contributions as percentages of endowment began at above 60 percent and rose to about 90 percent toward the end of the game.

Organizations can also prevent unraveling by providing an arena for interactions only among particular individuals, excluding those who are likely to deviate from the norm. Indeed, the sunk costs associated with entering and remaining in various religious groups, cults, communes, etc. may reflect exactly this function. Organizations, in this case, are means to **sort** individuals by their social preferences.

10.3 The Contingency of Social Preferences: A Situational Perspective

The premise in the above discussion has been that social preferences are attributes of individuals. Social preferences do not depend on the environment within which individuals

interact. Only their implications depend on it It has been thus implicitly assumed that just as we can take one's preference over goods or others' strategies as fixed in using micro-economic theory or game theory to study institutions, we can also take one's social preferences as given in studying behavior in various environments. While this assumption may be appropriate in various applications, it is also clearly partial. This section argues that social preferences are often situational. The manifestations of humans' capacity to have social preferences - what so far has been assumed to be parameters in one's social preference - depend on the situation.

The situational contingency of the manifestations of social preferences is well reflected in experiments. The design of the experimental studies discussed above was not aimed at examining the this situational contingency of social preferences and hence is biased against finding it. In these experiments, the set of relevant others was exogenously given, socially beneficial actions were clearly defined, and one's implications on the welfare of others were unambiguous. Furthermore, income to each player at the beginning of the experiment was pre-determined. Similarly, in the models of social preferences that were developed based on these experiments, various relevant parameters were taken as exogenous. The parameters of these models specify the relevant individuals, what is socially beneficial, what income should be divided with others and to what extent, the weight placed on the material welfare of others, and so on.

Even in this confined setting, experiments indicate that social preferences differ among societies. Henrich et al. (2001a, 2001b) is the most extensive experimental study, exposing whether there are systematic societal differences in the ways that social preferences express themselves. The study conducted various experiments in fifteen small-scale societies in twelve countries and five continents, exhibiting a wide variety of economic and cultural conditions. Three were foraging societies, six practiced slash-and-burn horticulture, four had nomadic herding groups, and three were sedentary, small-scale agriculturalist societies.

Comparing results in various games revealed very large and statistically significant variations in behavior across societies. In the Ultimatum Game, for example, mean offers ranged from about 26 to 58 percent of the total stake. In some societies the mean offer was as low as 26 percent, in others it had been 30 or 40 percent, while in yet others it had been between 40 and 50 percent. In two groups the mean offer was greater than 50 percent, implying that people, by and

large, offered more to others than they took for themselves. Similarly, the sample modes vary from 15 percent to 50 percent. These large variations suggest that the manifestations of social preferences are society-specific. Indeed, even comparisons between modern nations yielded systematic differences. (Roth et. al. 1991.) If the unselfish behavior reflected in these outcomes is due to social preferences, one has to recognize that these results reflect systematic inter-society differences in their manifestations.

Furthermore, experiments in Ultimatum and Dictator Games illustrate that individuals recognize the social preferences that guide behavior in their society. In particular, they recognize what behavior will be considered by others as offensive and thereby invoke retaliation. In various countries, individuals seem to have a shared notion of what is considered insulting enough to cause rejection.[53] Indeed, in developed economies such notions are so well shared that econometric analysis suggests that individuals make offers that will maximize their expected returns. (Henrich et al. 2001a, 2001b.)

The sources of such cross-society variations are not the focus of the current discussion as it relates to the origins of institutions. But in any case, Henrich et al. (2001a, 2001b) argued that the two variables together can explain about 68 percent of the differences (variance). The first is whether in that society economic production requires cooperation among non-kin; the second is how much people rely on market exchange in their daily lives. Group mean offers in the Ultimatum Game are positively and significantly correlated with these two variables. Hence, it seems that at least 32 percent of the cross-society differences cannot be accounted for by at least these two economic variable and have, at least so far, to be considered as reflecting "cultural" distinctions. But these cultural distinctions may be larger because, as we have seen in part II, the nature of economic activities and the extent of the market cannot be taken as exogenous to the beliefs that prevail in a society. I will return to this issue in part VI.

Experiments conducted in other settings also suggest that members of a society share and condition their behavior on particular, context-specific notions of fairness and equality. The parameters in individuals' social preferences depend on the situation. Hoffman et al. 1994

---

[53] Recall there are two ways to interpret these results. The first is that they reflect inequality aversion (e.g., Fehr and Schmidt 1999). The second is that they reflect outcomes in a psychological game (E.g., Charness and Rabin 2001). The discussion here is based on the second interpretation.

compared results in Ultimatum Games in the US in different settings. One was the regular setting in which the proposer was randomly selected. The other setting was one in which it was known to both players that the right to be the proposer in an Ultimatum or a Dictator Game was earned by a high score on a general knowledge quiz. The idea that being a proposer is a "right," was reinforced by the instructions for the experiment. The results between the two settings were statistically significant and indicate that people were willing to provide another with less if they earned the right to propose an allocation.

In Dictator Games, for example, when the proposer was randomly assigned, about 20 percent of them gave nothing to the other, and about 75 percent gave $3 or more to the other out of the $10. When the right to propose was assigned based on the quiz's results, more than 40 percent of the proposers gave nothing to the other and only about 20 percent gave $3 or more. The fact that one has "earned" the right to be the dictator influenced the perceptions of the interacting individuals regarding what a fair allocation was.

Empirical evidence also indicates that even contemporary societies have substantially different norms related to who the relevant others are, what is fair, and what is an appropriate norm of equality. Platteau (2000) has documented the large extent to which sharing norms prevails in sub-Sahara Africa. If one's harvest was particularly plentiful, it is considered a matter of luck and hence has to be divided among others. Needless to say, this is not the norm of fair allocation that prevails in US agro-business. But even in the US, there seem to be different norms of how to allocate the surplus from agricultural production. Young and Burke (2001) conducted a detailed analysis of agricultural contracts in contemporary Illinois and examined how the profit is divided among landlords and cultivators. Controlling for land quality and other economically relevant characteristics, they found that in northern Illinois the surplus is divided mainly equally, while in southern Illinois only 14 percent of the contracts have equal division. Instead, more than 50 percent of the contracts provides the cultivator two-thirds of the profit.

Even modern national economies seem to be organized around distinct norms of who is responsible for acting altruistically toward whom. In the contemporary Japanese welfare system, for example, family members were expected to contribute to the welfare of those among them who were unable to care for themselves. The Civic Code Article 877 specifies a legal obligation

to support family members within three lineal generations.[54] Under this Civic Code, one has some obligation to pay, for example, the living costs of a disabled family member. This is not the case in the US in which, legally at least, family members do not have this responsibility.

On the international level, individuals seem to consider mainly citizens of their own nation to be the "relevant others" toward whom they should act altruistically. Private, voluntary charity to the poor is confined mainly to members of the same nation as those who provide the charity. Historically, this has not been always the case. For example, the old English poor law regulated assistance to the poor for over two centuries. A parish was responsible for its poor but a parish could also support any poor it so desired. Yet, as is well known, parishes sent the poor who came to their doors back to the parish that was obliged to care for them even if it was able to provide less assistance.

The contingency of the manifestation of social preferences is well reflected in the failure to find a general model of social preferences. Andreoni and Miller concluded (2002: 20) that their effort to apply the model of social preferences they derived based on experiments in Dictator Games "suggests that many things other than the final allocation of money are likely to matter to subjects. Theories may need to include some variables from the game and the context in which the game is played if we are to understand the subtle influence on moral behavior like altruism." In surveying other specifications of social preferences, Fehr and Schmidt (2001) have reached similar conclusions.

On the other hand, an axiomatic approach for social preferences (Segal and Sober 2000) supports the claim regarding the primacy of a society's influence on one's preference. They have examined what axioms are required - have to be taken as exogenous to the analysis - to derive a utility function which includes social preferences, meaning one which is a weighted average of the material payoffs of the individual under consideration and those of the other players. To achieve this they had to take as axiomatic that each player has an ordering over his opponent's strategies that captures their "niceness." In other words, to derive an individual's utility function it has to be assumed that each member of the society has internalized a norm

[54] In April 2000, however, "the Care Insurance Act" established a kind of social insurance which is compulsory for those over forty years of age. Its purpose is to make the third party, such as a service company or public organization, supply the care-service.

regarding how nice the various strategies of others are toward him.

Such experimental, empirical, and theoretical results support the argument that although it may well be that social preferences reflect a universal human propensity, their manifestations in various societies over time and situations seem to be situational.

10.4 Social Preferences and Internalized Institutional Elements

The human capacity to have social preferences seems to be universal. Their manifestations, however, seem to differ among societies and even within societies across situations. Humans may generally have, for example, the capacity to care about the welfare of non-kin, but this general tendency is not what determines whether one cares about the welfare of his extended family, the village, the town, the nation, or the world. Humans may generally have the tendency to be reciprocators or to care about what others think of them, but this general propensity, in and of itself, does not account for why individuals in some societies feel bad for failing to pay a 15 percent tip to a helpful waiter but do not feel bad for not tipping a helpful cab driver. Our general tendencies do not explain why, for example, until the present day, the same behavior toward a teenager's girlfriend in the northern and southern US by another teenage boy is likely to provoke different emotional reactions.

This suggests that we have to complement the view of social preferences as fixed aspects of a utility function with another view. This complementary view considers social preferences themselves as endogenous. The premise that, for example, inequality aversion is a universal property of some members of a society is replaced by the premise that some individuals have the capacity to be inequality averse. The exact manifestation of this capacity in a particular time and place, in turn, reflects the social malleability of human preferences. The focus of the analysis is thus on the endogenous formation of what the above discussion has considered as parameters in one's social preferences. Similarly, while maintaining the premise that individuals are reciprocators, this view concentrates on the need to examine the determinants of what one considers as fair or unfair and the determinants of the behavior that one maintains that will cause others to resent him.

Such determinants of social preferences are endogenous to the society, exogenous to each individual, and influence behavior in recurrent interactions among individuals with particular

social positions. Hence, such determinants of social preferences are institutional elements. Indeed, they are internalized institutional elements in the sense that they influence behavior through the psychological and emotional benefits and costs of behaving or failing to act in a particular way because these behavioral standards were internalized by the individual and were integrated into his utility function.

Institutional analysis, for example, is about the beliefs that people bring to a particular interaction regarding the intentions associated with particular actions. When a driver raises a particular finger to another, the intention beyond the action is rather clear to both although it was not established through the interactions between these two particular drivers. But for such beliefs about intentions to have an effect, to influence one's well-being, they must have been internalized by this individual. It is the driver's inability to prevent his own negative emotional response that provide the other driver's with the possibility of influencing the way he drives by the expectations that certain actions would lead to this gesture.

Hence, the study of social preferences from a situational perspective inter-relates with a broader issue, that is, specifically, the study of internalized institutional elements: the influence of a society on its members' social preferences. We are far from having an analytical framework that will enable us to study such internalized institutional elements. Game theory, augmented by insights from various disciplines, however, contributes to such a framework.

10.4.1 Studying Internalized Norms

Game theory facilitates the study of social preferences through its more general contribution to the study of internalized norms. The notion of internalized norms should not be confused with two other, commonly used, and inter-related in reality notions of norms. The first notion is that of a social norm as a rule of behavior that is neither promulgated by an official source, such as a court or a legislature, nor enforced by the threat of legal sanctions, yet is regularly complied with. (Section 7.1 and chapter 8.2.2) The second notion of a norm is as a rule of behavior that specifies what is morally appropriate, good, true, and beautiful.[55]

The third notion of norm, and the one which is relevant here, is that of an internalized

[55] See, for example, Shapiro 1983: 25; Davis 1949: 52.

standard of behavior that is embodied in one's preferences. Parsons (1951: 38-40) has taken the position that full institutionalization of a behavioral standard requires its internalization. Internalization, the incorporation of behavioral standards into one's superego, essentially means the development of an internal system of sanctions that supports the same behavior as the external system.[56] Even in the absence of any external motivation, one who has internalized the norm of keeping a promise may avoid the temptation of material gains from reneging. Peoples' non-material, intrinsic utility from behaving in a particular way is determined by the extent to which they have internalized particular norms.[57]

Scholars from Durkheim to Elster have argued that internalized norms are essential determinants of behavior in societies; they are the cement of society (Elster 1989). All known societies foster norms that enhance personal fitness, such as prudence, personal hygiene, and control of emotions. Societies also universally promote norms that subordinate the individual to group welfare, fostering such behaviors as bravery, honesty, fairness, willingness to cooperate, refraining from over-exploiting common pool resources, contributing to political life, acting on behalf of the larger community, and identifying with the goals of the organization of which one is a member (Brown 1991).

It is easy to capture the effect of internalized, exogenously given norms in game-theoretic framework. Taking an action which is against the norms one has internalized reduces that person's utility, everything else being equal. In economics the motivation induced in this way is sometimes referred to, in the context of contract enforcement, as "first party enforcement." (E.g., Ellickson 1990; Greif 1997; Aoki 2001.[58])

Internalized norms also regulate the manifestations of social preferences. Unlike the laboratory setting in which the set of relevant others, the relevant income, and what a "fair"

---

[56] Regarding norms and their transmission, see, for example, Cavalli-Sforza and Feldman 1981; Bandura 1971; Witt 1986; Shapiro 1983.

[57] Intrinsic motivation is defined in psychology as the motivation to take an action despite the lack of any reward from doing so except for the value of the action itself. See review in Frey 1997: 13-4. See also Kreps 1997.

[58] Dawes and Thaler 1988 defined "impure altruism" as an action benefitting others while being motivated by internalized norms of behavior. One acts altruistically, motivated by the benefit of feeling you are doing the right thing.

distribution is, are, to a large extent, determined exogenously, in the real world such parameters in social preferences are internalized norms. A norm that one has internalized causes him to care about the welfare of certain others apart from his immediate kin. A norm determines whether they are members of his extended family, the village, the town, the nation, or the world. Similarly, how much one cares about the inequality of others reflects internalized norms. The equality norms that one has internalized determine whether he will gain utility from dividing his income equally among the "relevant others" and if not, how much to give.

Studying parameters in social preferences as institutional elements thus requires an analytical framework that restricts the set of arguments regarding admissible norms and exposes the mechanism through which norms generate behavior and are generated by it. Game theory, augmented by insights from various disciplines, contributes to developing such a framework.

Sociologists have long argued that the internalization of norms mainly occurs through a socialization process. (Durkheim 1951, Mead 1963, Parsons 1967). Socialization is carried out by parents (vertical transmission), by one's peers and role models (horizontal transmission), and socializing institutions (oblique transmission) such as schools and churches. Game theory provides an analytical framework to study some aspects of these three socialization mechanisms. This framework explicitly considers the inter-relationships between norms and behavior in a given environment and derives both internalized norms and behavior endogenously.[59]

Evolutionary game theory enables studying vertical transmission as reflecting economic competition among various individuals who have internalized various norms. A classical analysis is provided by Frank (1987).[60] He conceptualized the propagation of norms as reflecting the transmission of traits from parents to offspring. Parents with more economically successful traits are assumed to have more offspring. To capture how vertical transmission operates, it is explicitly modeled as a replicator dynamic. A replicator dynamic is a functional form expressing how, over time, there is a shift in the population toward more successful traits whether they are genes or behavioral patterns. (Weibull 1995.) We can interpret this as reflecting the fitness

---

[59] This section presents the analyses of the first two mechanisms while a particular aspect of the third is examined in section 10.6.

[60] But see Harrington 1989.

advantage of successful traits. Those who have adopted them have more offspring. Hence, the assumption is that the frequency of internalized norms positively responds to the material well-being they imply.

To examine how the norm of honesty can prevail in a society, Frank examined a Prisoner's Dilemma (PD) game. He noted that if one has strong feelings of guilt when breaking a promise, that person will often honor his promises even when material incentives favor breaking them. Such an individual can therefore commit to cooperate in the PD game and will find it materially beneficial to do so if the other player can commit as well. "It is precisely this capacity of emotional forces to override rational calculations that makes them candidates for commitment devices" (p. 594).

But such normative behavior will materially benefit one only to the extent that he can communicate that he has internalized this norm and it constrains his behavior. "Merely *having* a conscience does not solve the commitment problem; one's potential trading partners must also know about it... A strategically important emotion can be communicated credibly only if it is accompanied by a signal that is at least partially insulated from direct control" (p. 594). Blushing, sweating, and movement of the eyes can and do serve as such communication devices. When this signal is imperfect, the equilibrium implied by evolutionary forces will contain both honest and dishonest individuals. Some members of the society will internalize the norms of honesty. On the other hand, some people can, at some cost, pretend that they have internalized the value of honesty. If the population contains no honest people this behavior implies a lower payoff than simply cheating. There are no honest people to fool in any case. However, the value of adopting such behavior increases in the number of honest individuals who can be fooled. Hence, the long-term equilibrium of this system is likely to have a mix of honest and dishonest individuals.

The analysis enables exploring the relationships between various parameters and the frequency of honesty in the population. This frequency increases, for example, the higher the cost of pretending to be honest is and the lower the gain from cheating. Clearly, the model can also be expanded to explore such factors as organizations that further inflict punishment on cheaters.

Game-theoretic models of horizontal transmission are based on extending the basic

game-theoretic framework by specifying a utility function that depends on one's normative behavior in addition to the material payoffs. An example is Andreoni (1990) who developed a model of private benefit - "warm glow" - from adhering to the norm of contributing to the public good. This benefit reflects a private utility from impure altruism, namely, doing the morally appropriate thing. Brekke et al. (2002) presented a version of this model in which the strength of the warm glow depends on the difference between what one considers the morally appropriate behavior to be - the socially optimal behavior - and the actual behavior. The morally optimal behavior, in turn, is determined by each individual based on his knowledge of the situation.

It is easy to see how such game-theoretical models can be extended to capture horizontal transmission: one's normative behavior is a function of the behavior of others. In particular, such an extension can be made along the lines developed in Lindbeck (1997) and Lindbeck et al. (1999).[61] In this formulation the existence of a norm is taken as exogenous. The norm may be, for example, that of working for a living. Its intensity, as felt by each individual, however, is endogenous: it depends on the number of people adhering to it. When the population's share of welfare recipients who do not work is large (small), an individuals's discomfort from such a lifestyle is relatively weak (strong). Hence, the intensity of the norm can be examined as an equilibrium phenomenon and the equilibrium level of a norm in a particular environment can be explored.[62]

??? A simple model of a strategic version of this analysis will be added. The essence of the model would be that utility from taking a particular action increases with the number of people expected to take it.

The above discussion took the environment within which individuals interact and in which internalized norms establish themselves as given. But similar to previous discussions, the

---

[61] Their framework is non-strategic.

[62] See also Akerlof and Kranton 2000 who examined the implications of a situation in which students can make an irreversible choice from a fixed menu of identities (bundle of values). While they take the menu of identities as given, the analysis can be extended along the above lines to examine the distribution of equilibrium identities.

analysis can be, and for a comprehensive understanding of an institution has to be, further extended to consider other institutional elements such as behavioral cultural beliefs and organizations. In this regard, it is worth noting the additional role of organizations here in preventing the unraveling of normative behavior.

The power of a norm to command adherence is positively associated with the level of adherence in the population. The more others adhere to a norm, the higher one's non-material benefit from adhering to it will be. This implies that a norm can unravel when different people have distinct non-material benefits and material costs from adhering to it. If this is the case, after some people cease following the norm, some others may also cease following it as well because a reduction in the number of those who follow the norm reduces the benefit of doing so. This may lead still others to do the same, leading to the continuing unraveling of the norm.

Organizations can **prevent** such **unraveling** of normative behavior in various ways. First, by punishing the first ones to deviate from it. Second, organizations influence such unraveling through their impact on the relationships between intended normative behavior and actual outcomes. Normative behavior can unravel because individuals who adhere to it observe that their actions are in vain. If one's donated blood is sold instead of given away, one is less likely to donate blood, thereby leading to the unraveling of the norm. Enforcement against selling blood by the police or hospitals may thus be crucial for the perpetuation of the norm.[63]

In sum: the particular norms regarding behavior, relevant others, fairness, and equality that were internalized by members of a society and influence their behavioral choices in various situations are institutional elements. They influence behavior through the associated psychological rewards and costs. In the case of social preferences, the norms that were internalized determine the psychological costs and benefits of taking an action influencing another person's welfare. These norms are beyond the control of each of society's members who internalized them, but they are nevertheless generated endogenously. As we have seen, game theory contributes to the study of such regeneration by highlighting the links between behavior,

[63] Organizations can also limit the extent of free-riding on internalized institutional elements. Lindbeck and Weibull 1988 illustrates that a norm of assisting others may be undermined and lead to inefficiencies by the implied moral hazard problem. One's altruistic inclination reduces the incentive to others to engage in wealth-creating activities. Instead, they can rely on the other's support. Organizations such as a compulsory social security system can mitigate such problems.

norms, and outcomes.

10.4.2 Studying Emotions using Psychological Games

To study the manifestations of reciprocity as an institutional element note its relationship with emotional responses. In studying emotions, one can differentiate between three issues, the first being the origin of the capacity to have an emotion. Section 10.4.1 touched upon this issue when presenting Frank's (1987) model in which emotional responses associated with honesty evolved. The second issue is the analysis of non-psychological emotions that do not depend on the actions on others, such as malice or hate. Such emotions can be integrated into the rules of the game and their implications on admissible behavioral cultural beliefs can then be examined using equilibrium analysis.

The third issue is that of psychological emotions, namely, uncontrolled emotional responses such as gratitude, anger, joy, shame, and guilt, that depend on or are triggered by, another person's actions. Such emotions are directly relevant to the study of the manifestations of the propensity to exhibit reciprocity. Reciprocity manifests itself when one has an emotional response to an action taken by another. In particular societies, individuals in general will respond emotionally if another driver shows them a particular finger. In other societies, an elderly women would be emotionally offended if a young man did not stand up when she entered the room. In some societies, a woman who is paid less than a man for the same work would be emotionally offended, while in others that would not be the case.

The particular emotional responses of others that establish themselves in a society and are internalized by many are institutional elements. Although endogenous to the society, they are exogenous to each of the interacting individuals who have internalized them and whose behavior they influence. The analytical framework provided by game theory enables deductively restricting arguments regarding the nature and magnitude of such emotional responses in a given environment. Furthermore, game theory provides an analytical framework for capturing mechanisms, causing them to generate behavior and to be generated by it.

Game theory facilitates the study of emotional responses to various actions through psychological game theory, which was developed by Geanakoplos, Pearce and Stacchetti 1989

(henceforth GPS).[64] It provides an analytical framework for examining the endogenous manifestation and economic implications of emotions.[65]

Recall that in game theory one's utility depends on the strategy choices of all the other players. Hence, one's utility *indirectly* depends on beliefs about such choices through their influence on strategy choices. In psychological games, at least one player's utility also depends *directly* on his beliefs about another player's choices and, possibly, beliefs about such beliefs about choices, and so forth. In general, "the players' payoffs depend not only on what everybody does but also on what everybody thinks" (GPS: 61).

Given this setting, psychological game theory requires that in equilibrium, beliefs about behavior and actual behavior coincide. Emotional responses to one's actions, therefore, are not exogenously fixed. They depend on the equilibrium beliefs regarding actions. If, for example, the equilibrium play is for one to be honest but that person nevertheless cheated, the other may have an emotional response of anger which, in turn, may imply that it is optimal for him to punish the cheater despite the implied material costs. Emotions are determined endogenously in equilibrium and those that do not express themselves on the equilibrium path, such an anger or shame, can nevertheless have a profound influence on outcomes.

Psychological games are rather complex and hence it is best to provide a simple example illustrating the basic idea. The example is an extended Game of Trust. (Section 4.2) In this game, player 1 can either Trust (initiate exchange) or not, and if he does not, both players get the payoff of zero. If he chooses trust, player 2 can either be honest or cheat. Denote by p the probability that he is honest and by 1 - p the probability that he cheats. If he is honest, each player gets the payoff of 3. If player 2 cheats, however, and in contrast to the original game of trust, player 1 can either take costly revenge or not. If he does not take revenge, his material payoff is 1. Revenge, however, is materially costly, implying that if player 2 takes it, his payoff

---

[64] See also Koplin 1992 and the reference below. For applications see, for example, Huang and Wu 1992.

[65] The analysis by Holländer 1990 presented in section 8.1 captures how classical game theory also facilitates the study of emotion. Each individual feels emotionally uncomfortable when taking actions different from those of the group in general, and the actions are costly to him. In equilibrium both the levels of emotion and behavior are determined.

is only .5. Player 2's payoff is 5 if no revenge was taken and 1 if it was. The game is presented in the figure below. The "r" in player 1's payoff captures his non-material payoff from taking revenge. Ignore this payoff for the moment.

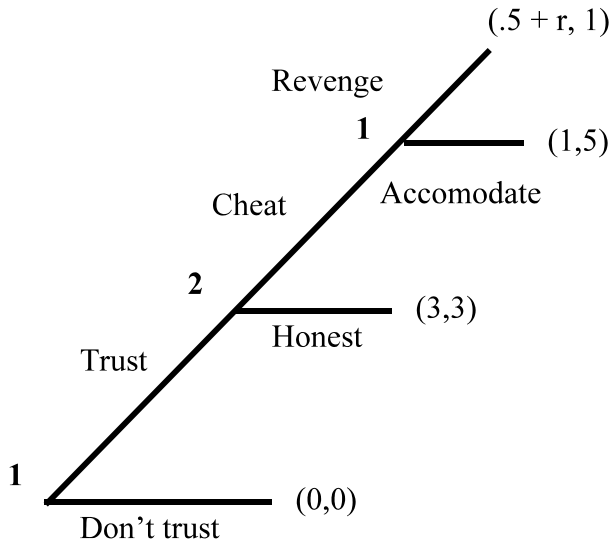


Figure 10.1: The Game of Trust: Taking Revenge

Ignoring the non-material payoff and conducting a backward induction analysis, it yields that in the unique equilibrium, player 1 does not trust player 2. Taking revenge is materially costly implying that the theat of doing so is not credible. Hence, player 2 will cheat and player 1, expecting this to be the case, will not trust to begin with.

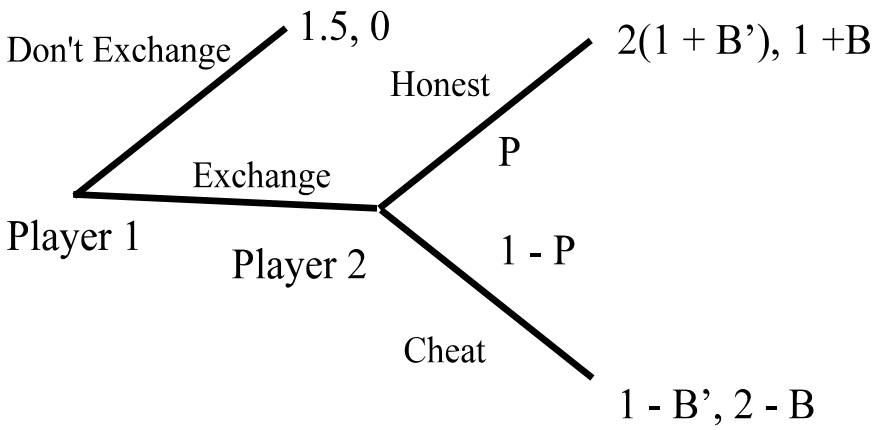
Now consider the same game with emotions. Specifically, suppose that player 1 gets the emotional payoff of r from taking revenge if he took it after he was cheated. The size of this payoff, however, is not taken as exogenous to the analysis. It is determined as part of the equilibrium. In particular, the emotional response is proportional to deviation from the expected behavior of player 2. If cheating is expected, $p = 0$, then revenging is lame, $r = 0$. Revenging only reinforces the feeling of being in a position in which a reasonable person should not be in to begin with. But if honesty is expected, $p = 1$, then revenging is sweet, $r = 1$. The equilibrium analysis imposes the restriction that beliefs about the behavior of player 2, namely p, are equal to the beliefs in emotional response, r.

In the game with emotion there are two equilibria. In the first, the strategy is {Trust, Honest, Revenge} and the beliefs are $p = r = 0$. In the other equilibrium, the strategy is {Don't

Trust, Cheat, Accommodate} and the beliefs are p = r = 1. In this latter case, both players are better off. The threat of revenge is credible but it is off the path of play, implying that it is not exercised and leaves both players better off.

To illustrate a somewhat more complicated analysis of psychological games, consider again the game of trust. In this specification of the game, if player 1 doesn't trust, he gets 1.5 and player 2 gets nothing. Ignoring emotional payoffs, if trade was initiated, player 2 can gain 2 from cheating and 1 from being honest. Due to emotional response, however, player 2's payoff depends not only on what he does but also on what he thinks player 1 thinks of his character. Denote by B the expectation that player 2 holds regarding what player 1 thinks of him. In particular, suppose that player 2 gets satisfaction - pride - from the thought that 1 thinks he is honest and regrets, feels shame, when this is not the case. Player 1 cares about what player 2 does but also takes emotional pleasure from believing that 2 is honest. Hence, let B' represents player's 1 belief regarding player 2's action. Denote by p the probability that, on the equilibrium path, player 2 chooses to be honest. In equilibrium it must be the case that p = B = B'. This game is presented below.

Figure 10.2: The Game of Trust: Guilt and Emotional Satisfaction



In the absence of emotions, player 2's best response is to cheat. But in the game with emotions, this is the case only if it is expected that 2 will cheat, that is, p = B = B'= 0, implying that the game is the regular trust game. Its interpretation, however, is different: If individuals in player 2's situation are expecting to be cheated, cheating does not imply guilt. Hence, player 2 will still get 2 from cheating and 1 from being honest, making cheating his best response. Trust would not be initiated to begin with. But there is also an equilibrium with trust and honesty. If both players share the expectation that individuals in player 2's situation will be honest, that is p = B = B'= 1, then player 2 will feel guilt from cheating and pride from being honest, implying the payoff of 2 from honest behavior and 1 from cheating. Hence, player 1 can trust to begin with.

Psychological game theory thus captures how deviation from expected equilibrium behavior can trigger an emotional response. This emotional response, in turn, can support an equilibrium which otherwise could not have been sustained although the emotional response is off the equilibrium path. In classical game theory, the expectation that one will respond to another's action by taking an action that reduces his material payoff is not credible. Hence, in games of complete information, the threat of taking it cannot support an equilibrium behavior. Psychological games, however, capture that emotional responses can render such actions credible.

Psychological game theory has been extended to provide an analytical framework within which we can deductively examine reciprocity. Reciprocity reflects one's emotional response, gratitude or hostility to what is considered to be fair or unfair behavior by another. The basic idea was formulated by Rabin (1993).[66] He postulates that people want to be nice to those who treat them fairly and want to punish those who hurt them. An action is perceived as fair if the intention that is behind it is kind, and as unfair if the intention is hostile. The kindness or the

[66] Rabin 1993 examines two players' normal-form games. Dufwenberg and Kirchsteiger 1998 generalized the theory to N-person extensive-form games and introduced the notion of a "sequential reciprocity equilibrium." Their specification captures how beliefs are formed off the path of play and hence they can define an equivalence to subgame-perfect equilibrium. Falk and Fischbacher 2000 extended the framework for the case of incomplete information although they assumed that player1's action is perceived as kind by player 2 if it implies a higher payoff to 2 than to 1. Hence, their framework combines equality considerations with intentions. Charness and Rabin 2001 provide a similar extension.

hostility of the intention, in turn, depends on whether the payoff distribution induced by the action is equitable. Thus, the model is based on the notion of an equitable outcome but explicitly captures the role of intentions.

More specifically, Rabin assumes that each player's subjective expected utility (which includes material and psychological payoffs) depends on three factors: the player's strategy, his beliefs about the other player's strategy choice (first-order beliefs), and beliefs about the other player's beliefs about his strategy (second-order beliefs). A "kindness function" is a measure of how kind player 1 is to player 2 and it depends on the action that player 1 took and player1's beliefs about what strategy player 2 is choosing. If player 1 believes that player 2 is choosing a particular strategy, how kind is player 1 being by choosing that strategy? Given his beliefs about what player 2 is doing, his choice of action determines the possible payoffs for player 2. Hence, it is possible to obtain a kindness function: How kind was player 1 in choosing a particular action given his beliefs about the actions of player 2? It measures how "fair" player 1 was to player 2, given his beliefs. A similar kindness function can be defined for player 2.

The implied kindness functions are used to specify a subjective utility function for each player that captures the payoff from actions and the payoff from the perceived kindness or fairness of the other player. If the other is perceived to have behaved fairly, one's utility is increasing from reciprocating but the opposite holds if the other is perceived to have behaved unfairly. A game with such preferences is a psychological game and we can analyze it accordingly while imposing the restriction that in a fairness equilibrium, actions and beliefs are consistent with each other. The expected behavior is the same as the actual behavior. An interesting feature of this formulation is that it captures the empirically relevant trade-off between fairness and material considerations (Rabin 1993: 1284). As material payoffs increase, the impact of fairness decreases.[67]

---

[67] Roughly speaking, the set of fairness equilibria converges to the set of Nash equilibria as material payoffs increase. In further support of the point made here regarding the social nature of the criteria used to evaluate others' intentions, note that the most extensive model of fairness equilibria (Charness and Rabin 2001) based on psychological games begins the analysis with an exogenously given "selfless standard" and derives the equilibrium "demerit function." How much does player i deserve from player j's perspective if i took a particular action? But this merit function is not specified outside the equilibrium.

In sum, the emotional responses triggered by various actions that established themselves in a society are internalized institutional elements. When two individuals interact, they bring emotion to the interaction. They share a view regarding what action will lead to one emotional response or another. One paid a 15 percent tip and feels good afterwards. One who did not let another to take his turn entering an intersection knows that the other will be angry at him. Once particular emotional responses have established themselves in a society, they are exogenous to each individual, influence behavior, and are regenerated by this behavior. Psychological game theory enables us to study exactly this. It deductively restricts arguments regarding admissible emotions and exposes inter-relationships between aggregate behavior and individuals' emotions.

10.5 Rules and Communication

In studying behavioral cultural beliefs we noted the role of publically articulated and commonly known rules in making a situation common knowledge, defining various states in it, and coordinating on a particular behavior. Rules serve this role with respect to internalized institutional elements as well. In addition, the game-theoretic framework enables considering the distinct nature of the coordinating role of rules when internalized institutional elements are endogenously determined. To see this, recall the role of rules in coordinating behavior when internalized institutional elements are ignored. In this case, a rule coordinates on one among the many possible self-enforcing regularities of behavior.

Now consider a game-theoretic analysis of, for simplicity, internalized norms. Recall that the intensity of a particular norm depends on the number of people who follow and are expected to follow it. Hence, a rule that influences peoples' expectations about the behavior of others also changes their motivation of whether or not to follow it by altering the associated payoffs. Following the rule that others follow implies both receiving the material payoffs associated with it as well as the non-material, normative payoffs. Hence, behavior that would not be self-enforcing if norms were not being considered, can become self-enforcing.

This role of rules in coordinating behavior that becomes normatively self-enforcing is reflected in experiments. Ostrom (1998: 7) has noted that when people can communicate and agree on rules of behavior, they seem to follow the behavior they have agreed upon even if it is not in their best interests. "Subjects in experiments do try to extract mutual commitment from

one another to follow the strategy they have identified as leading to their best joint outcomes. They frequently go around the group and ask each person to promise the others that they will follow the joint strategy. Discussion sessions frequently end with such comments as: Now remember everyone that we all do much better if we follow X strategy." When such discussions were allowed in experiments, cooperation levels increased.

Similarly, empirical studies also reflect the role of rules in coordinating behavior that becomes normatively self-enforcing. Stewart (1992), for example, has noted that in relatively similar economies such as the US and UK different rules regulate the donation and selling of human blood. He found that members of societies whose legal rules preclude the sale of human blood for medical purposes but encourage donations, hold stronger norms against selling it.

Rules that are able to alter expected behavior and feedback into norms can be referred to as normative rules. The only restrictions that game theory imposes on the set of such rules is that the behavior they specify has to be an equilibrium given its normative impact. This begs a central question: What determines which rules will influence expected and actual behavior to create this normative impact? This question touches upon the origins of institutions that is discussed in later chapters. But it can be noted that although game theory highlights the importance of this question, it sheds little light on how to address it.

In any case, it seems that so far we cannot escape explaining why a particular rule gains the power to influence norms, not just behavior, by invoking another ill-specified concept. The concept is that of the legitimacy of the process through which the rule was articulated. This legitimacy, in turn, depends on the prevailing norms in the society under consideration. In democratic countries, for example, it seems that having a voice, participating in the decision process through which the rule is specified, confers legitimacy on it.

Frey (1997, chapter 7), for example, has provided econometric tests of the relationships between constitutional rules and civic virtue. He found that tax compliance in Switzerland differs among cantons. Many of these differences, in turn, are explained by the different degrees of political participation possibilities. When people participate in determining rules, they follow them more than can be explained by only the private benefits and costs of tax evasion.

10.6 The Evolutionary Origin of Prosocial Preferences

Social preferences and emotions that are beneficial to the society but are harmful to those who carry them present a puzzle of prosociability. Why have people developed preferences for enhancing social welfare despite the fact that they also reduce their material benefits? (In this section I will refer to such preference as prosocial or altruistic.) As we have seen in 10.4.1, there is a long tradition in sociology arguing that such behavior reflects the socialization process. Socialization by parents (vertical transmission), by socializing organizations such as schools (oblique transmission), and by peers (horizontal transmission) causes individuals to internalize particular norms. Once internalized, the "oversocialized" individual carrying these norms will follow them regardless of their consequences.

This theory, that does not place a limit on the ability of socialization to influence behavior, has been criticized even in sociology (Wrong 1999). Evolutionary biologists accept the importance of socialization but suggest limiting its power to reflect evolutionary selection processes. Only norms that pass the test of biological fitness can survive. Only fitness-enhancing norms can establish themselves in the population and their evolution is based on Darwinian selection (Cavalli-Sforza and Feldman 1981, Lumsden and Wilson 1981, Boyd and Richerson 1985). Indeed, in discussing the work of Frank (1987), we have already seen how such an argument is applied to analyzing the evolution of the norm of honesty.

It is difficult, however, to advance a similar evolutionary argument regarding social preferences that seems harmful to those who hold them, such as altruism. After all, those who act altruistically, help others, and contribute to the public good are likely to get a lower payoff than others who do not. Hence, altruistic individuals are likely to have a lower material reward and therefore will be less likely to succeed in the evolutionary competition with others. Altruism among strangers (interpreted as a Tit-for-Tat behavioral strategy) has been explained as reflecting evolutionary forces by Trivers 1971 and Axelrod and Hamilton 1981. They state that it can be sustained under evolutionary competition when interactions are repeated indefinitely among individuals with selfish, materialistic preferences. It is the prospect of future gain under a credible threat of future punishment that induces sufficiently patient selfish individuals to make material sacrifices in repeated games. But this a concept of altruism cannot account for the findings that individuals are altruistic in one-shot interactions. Wilson (1975) has identified the explanation of non-reciprocal altruistic behavior as a central problem in evolutionary biology.

Similarly, evolutionary economists consider behavioral patterns as reflecting strategies that emerge through an evolutionary process. The emergence of individual strategies is assumed to reflect a replicator dynamic process. Recall that in this process the number of individuals using strategies yielding a higher payoff increases over time. Hence, because altruism imposes a cost on its carrier, it cannot survive in the long run. In economics, as we have seen, it has been credited with reflecting repeated interactions among selfish individuals. (Chapter 4.)

The lack of either convincing sociological or evolutionary explanations for altruistic behavior has lent support to economists' inclinations to adopt a model of selfish individuals, but experimental evidence indicates its limitations. More recently, various explanations for prosocial preferences have been advanced. Presenting the details of these explanations is beyond the scope of this work but I will briefly describe some of them. This is done to highlight that a main argument that this work advances is at their core: the interplay between individuals' actions and the social factors behind the control of each of them (particularly organizations) is central to understanding socially determined outcomes.[68]

Field (2001, chapter 2) has built on recent work by Sober and Wilson (1998) and Wilson and Sober (1994) to argue that altruism can emerge from an evolutionary process when the selection process take place on both the individual and the group levels. When evolutionary selection is also done on the group level, social preferences may nevertheless be evolutionarily viable. Within each group, altruists will lose out in the competition for resources and, in particular, in the competition to pass on their genes to the next generation. Consequently, their share in each group will generally diminish over time. But because the behavior of altruists differentially benefits groups in which their frequency may be relatively higher, the proportion of altruists in the general population may rise in cases where the forces of group selection are stronger than those of individual selection. For altruists to not die out within each group, however, periodic recombining of, or migration among, groups is necessary. Altruists from the faster growing, more altruistic groups must periodically disperse throughout the global population. This ensures that their number in the population as a whole increases.[69]

---

[68] See also a review of other explanations in Field 2001, chapters 3-4.

[69] For a simple proof, see Gintis 2000, section 11.7.

There is now a large and evolving body of literature examining the evolution of altruistic preferences. The literature on preference evolution in games was pioneered by Güth and Yaari (1992) who argued that revengeful individuals can have an evolutionary advantage because no one wants to cheat them. This line of analysis has been further developed by Güth (1992), Bowles and Gintis (1998), and Huck and Oechssler (1999) among others. In each of these papers, the definition of reciprocal preferences is itself tailored to the specific environment under consideration. For instance, Guth and Yaari allow for individuals who have a preference for rejecting unfair offers in bargaining games, while Bowles and Gintis consider individuals with a taste for punishing free-riders in a model of team production. In contrast, Bester and Güth (1998) and Kockesen et al. (2000a, 2000b) have considered more general specifications of preferences that are defined independently of particular strategic environments, and depend only on the distribution of material payoffs in the group. Bester and Güth deal with the survival of altruistic preferences under pair-wise random matching, and Kokesen et al. with the survival of envious or spiteful preferences. Ely and Yilankaya (1997) and Dekel et al (1998) examine general models of preference evolution in which the class of preferences is composed of all possible orderings over action profiles.

To illustrate the flavor of this line of research and how, through interactions with experimental works, it facilitates advancing toward a better specification of preferences, consider the work by Sethi and Somanathan (2001). Their starting point is Levine's (1998) specification of altruistic preferences as depending on one's own material reward as well as those of others. That is, $U_i(x) = \pi_i(x) + \sum \beta_{ij} \pi_j(x)$, where x is the vector material payoffs, the summation is over $j \neq i$, and β_{ij} captures i's altruistic inclinations toward j. Specifically, Levine postulated that $\beta_{ij} = (\alpha_i + \lambda_i \alpha_j)/(1 + \lambda_i)$ where $-1 < \alpha_i < 1$ and $\lambda \geq 0$. The interpretation is that α_i is a measure of an individual's pure altruism, while λ is a measure of the degree to which the weight β_{ij} placed by individual i on the material payoffs of individual j is sensitive to the altruism of the latter. Levine demonstrated that this preference specification explains well the various results in experiments in such games as the Ultimatum Game and public good games.

Sethi and Somanathan pointed out that in an evolutionary competition, such preferences would be driven to extinction in various strategic environments. Accordingly they offered a modification of the above specification. Specifically, they suggested considering the case in

which $\beta_{ij} = (\alpha_i + \lambda_i(\alpha_j - \alpha_i)/(1 + \lambda_i)$. The only difference in specification is therefore that the weight that player i places on the well- being of player j not only depends on player j's measure of altruism, α_j, but on the deviation of player j's altruism from that of player i. In this case, an individual is altruistic toward those who are similarly inclined but is also capable of being spiteful and acts to reduce the welfare of others who are less altruistic than he is. This is the case when $\alpha > 0$ and $\lambda > 0$ and player i benefits from reducing player j's welfare when $\lambda_i(\alpha_j - \alpha_i) < 0$. This specification implies that individuals can be flexible in considering whether or not to be altruistic. Such reciprocal preferences are viable in evolutionary competition with purely self-interested preferences under various conditions.

Gintis has advanced another evolutionary process that may lead to social preferences. The interesting aspect of this analysis is that it explicitly captures the possible inter-play between the evolution of social preferences and the nature of existing organizations, such as social structures and socializing institutions. Gintis (2000, section 11.8) examined reciprocity, particularly the inclination to revenge inappropriate behavior. Even if this social preference implies an evolutionary disadvantage to the one who has it, it may nevertheless be evolutionarily viable because of interactions within one's group; it can provide fitness advantages. Social groups with an above-average share of such reciprocal individuals are better able to survive events such as wars, pestilence and famines that threaten the whole group with extinction or dispersal. In these situations, cooperation among self-interested individuals breaks down because future interactions among group members are highly improbable. But the presence of individuals who will retaliate for failing to cooperate will discipline the self-interested individuals so that the group is much more likely to survive.

Gintis (2001b) enriched the above purely evolutionary considerations by also considering the importance of socialization in studying how the interaction between evolutionary forces and socialization by such organizations as schools or churches leads to altruistic behavior.[70] The genotype - the genetic trait - of being able to internalize a norm can reduce its carrier's evolutionary fitness because of the implied psychological and cognitive prerequisites for the

[70] A similar earlier model is presented at Feldman et al. 1985. See also Gintis 2001a for similar analysis without the distinction between genotype and phenotype.

capacity to internalize norms. It can nevertheless be evolutionarily viable because it enables its carrier to internalize a selfish norm (a phenotype). This selfish norm is fitness-enhancing in the sense that it implies a higher reproduction rate. Gintis's explicit evolutionary model substantiates that this is indeed the case. The model assumes random matching between parents with particular genotypes, which they transfer to their offspring and thus transmit the norm through socialization by the parents. If the fitness advantages of the capacity to have the fitness-enhancing norm is sufficiently large, the evolving system of genotypes and phenotypes can reach a globally stable equilibrium of individuals with the capacity to internalize norms and have the fitness-enhancing norm.

To examine the evolutionary stability of an altruistic norm, Gintis extends the model to include norms which are fitness-reducing. As before, only individuals with the ability to internalize norms can be socialized into either or both norms. An altruistic norm, however, is fitness-reducing to its carrier but contributes to the society. He finds that if socialization is done only through parents, the altruistic norm cannot establish itself in the society. (It is not stable.) This is the case, in particular, because individuals with only the fitness-enhancing norm have a fitness advantage over those who have the fitness-reducing norm. Hence, for an equilibrium with altruistic norms to be stable, socialization from organizations other than the parents is needed.[71] Oblique transmission of altruistic norms through such socialization agents causes some individuals with the ability to internalize norms to switch to the altruistic norm. Thus, they enable the altruistic norm to survive. But because altruistic norms reduce the fitness of those who carry it, it cannot survive if transmitted only by parents; it depends on the fitness-enhancing norm to maintain the pool of individuals with the ability to internalize it.[72]

Hence, the possibility of internalizing an altruistic norm, the genotype required for internalization is due to the fitness-enhancing feature of the selfish norm. This selfish norm, as a phenotype, can establish itself in the society because over time there are more parents who transmit it to their offspring. This, however, is not the case for the altruistic norm whose carriers

---

[71] This ignores the issue of the motivation and financing of these socializing organizations.

[72] The result is weakened by introducing a replicator dynamic capturing horizontal transmission: A fraction of the individuals with the ability to internalize norms switches to the fitness-enhancing norm even if their parents do not have it.

are not likely to survive. As argued by Simon (1990) who inspired this analysis, altruistic norms, which are by definition fitness-reducing to their carriers, could 'hitchhike' on the general tendency of the internalization of norms to be fitness-enhancing. Finally, Gintis conjectured that the prevalence of pro- over anti-social norms in societies is due to the ability of groups with pro-social norms to win over groups with anti-social norms.[73]

Looking Ahead

Game theory, enriched by insights from other disciplines, provides an analytical framework to study various aspects of the inter-relationships between internalized institutional elements, behavioral cultural beliefs, organizations, and rules. This framework is still tentative at best, and its usefulness for positive, empirical institutional analysis is still uncertain, but its contribution is clear. Experiments based on game theory have enabled testifying to the importance of social preferences, paved the way to their exact formulation, and provided a way to study their impact on behavioral cultural beliefs and the broader context of internalized institutional elements.

Yet, it is also clear that advancing the above line of research is fraught with difficulties for which game theory at least has no satisfactory answers. The results of even well-designed experiments are often open to multiple interpretations. Analytically, we have a limited ability to model the trade-off between material and non-material rewards, or why in some situations social interactions lead to friendship and in others to animosity. Furthermore, in real-world situations multiple social preferences, norms, and emotions provide motivation for an individual. One may not act on his anger toward an employer who lowered his wage, knowing that the results will lower the welfare of his family members toward whom he wants to act altruistically. Similarly, we know that the "framing" of a situation will influence the way individuals will behave within it. One's framing of his own action, in one way or another, can thus be used to avoid the psychological costs of acting against one's norms.

[73] For other evolutionary models, see Sethi 2001, and see the literature review presented there.