Modeling Cooperation with Self-Regarding Actors: A Critique

Herbert Gintis

From a manuscript in preparation: Samuel Bowles and Herbert Gintis, **A Cooperative Species: Human Reciprocity and its Evolution**.

3

Relatedness, Repetition, and Reputation

> Because [the crocodile] spends its life in water, its mouth is filled with leeches. With the exception of the sandpiper, all other birds and animals run away from it. The sandpiper, however, is on good terms with it, because it is of use to the crocodile. When the crocodile climbs out of the water…the sandpiper slips into its mouth and swallows the leeches. This does the crocodile good and gives it pleasure, so it does not harm the sandpiper.
>
> Herodotus, *The Histories (1998) p. 122*

> Like successful Chicago gangsters, our genes have survived, in some cases for millions of years, in a highly competitive world. This entitles us to expect certain qualities in our genes. I shall argue that a predominant quality to be expected in a successful gene is ruthless selfishness. This gene for selfishness will usually give rise to selfishness in individual behavior.
>
> Richard Dawkins, *The Selfish Gene (1989) p. 2*

3.1 Introduction

Since Bernard de Mandeville (1924[1705]) and Adam Smith (1937[1776]), students of social life have stressed that cooperation sometimes occurs when it is in the personal interest of self-regarding individuals to contribute to the collective good. Adam Smith's "invisible hand" metaphor dramatized that private property and market competition can sometimes transform selfish motives into valued social outcomes. Thus, some forms of human cooperation result from institutional structures that make cooperating a best response for self-regarding individuals.

In human societies, we now know that Smith's invisible hand works if the contracts regulating all transactions are enforceable at no (or little) cost and cover all aspects of the exchange, so that there are no so-called "external effects" (Arrow 1971). In this case, economic actors take account of the effects of their actions on others for the simple reason that these effects are written into the contract: a benefit conferred by A on B is accompanied

by a compensating payment by B to A, and a cost imposed by A on B is accompanied by an equalizing transfer from A to B. Where external effects do arise, governments may intervene and directly enforce cooperation by punishing uncooperative acts. Even without the aid of contracts and governments, repeated interactions may sustain cooperation among self-regarding actors, with either the anticipation of retaliation or the possibility of loss of reputation deterring an infraction of the cooperative norm.

These are all forms of what we have termed, following biological usage, *mutualistic cooperation*. Biologists would add that cooperation results from the tendency of humans to interact frequently with those with whom they are genetically related, a predisposition to cooperate thus evolving by means of what is termed *kin selection*.

The dominant view in economics and biology for the past half-century has been that virtually all forms of cooperation among unrelated individuals are mutualistic. Biological theory has relied upon Robert Trivers' (1971) concept of *reciprocal altruism*. This theory assumes individuals help others only if their costs are, on average, fully compensated by the recipients. Trivers' idea was formalized by Axelrod and Hamilton (1981) as a repeated prisoner's dilemma, and was enriched by R. D. Alexander's (1987) notion of *indirect reciprocity*, in which individuals who are "in good standing" in the community cooperate with others who are in good standing in the community. If an individual fails to cooperate with someone who is in good standing, he falls into "bad standing," and individuals in good standing will not cooperate with him.[1]

The models of cooperation favored by economists also use a repeated game framework in which the threat of punishment, or of terminating a jointly profitable relationship, induces self-regarding individuals to sacrifice in the short run to sustain long-term benefits. Shubik (1959) suggested that game repetition and the threat of retaliation against defectors could sustain a high level of cooperation. The first formal statement of this principle was Friedman (1971), a special case of which is Trivers' reciprocal altruism. Fudenberg and Maskin (1986) provided a definitive model albeit, as we will see, under implausible assumptions concerning the information available to the players.

[1]Models of self-interested cooperation based on the good standing/bad standing distinction have been developed by Sugden (1986), Boyd (1989), Nowak and Sigmund (1993), Nowak and Sigmund (1998), Panchanathan and Boyd (2003) and others.

Repetition and reputation formation are doubtless important in fostering cooperation in human relationships where individuals often form long-term reciprocal bonds for mutual aid (Gouldner 1960, Homans 1961, Blau 1964). However, we believe many observed types of cooperation involving imperfect and private information concerning individual cooperative behavior, cannot be explained by repetition and reputation formation. This chapter, which deals with biological models, and the next, which deals with economic models, are devoted to substantiating this position.

We begin by reviewing the biological models of cooperation based on kin altruism.

3.2 Kin Altruism

Hamilton's rule (Hamilton 1964) for the evolutionary success of kin altruism shows that in a one-shot (non-repeated) interaction, conferring a fitness benefit b on another individual at a fitness cost to oneself of c will be favored by natural selection if $r > c/b$, where r is the genetic relatedness between the actor and the beneficiary. The early twentieth century British geneticist, J. B. S. Haldane anticipated Hamilton's reasoning when, asked if he would jump in the river to save his drowning brother, he reportedly responded "No. But I would to save two brothers or eight cousins."

When cooperation is favored by Hamilton's rule, it is altruistic, but it is easily explained, as the altruistic act in fact increases the frequency of the altruistic gene; i.e., the *inclusive fitness* of the actor is enhanced (by an amount $rb - c$). Kin altruism is widely observed in the care of offspring in many animals, and among humans, additionally, in such diverse aspects of behavior as food-sharing among adults (**?**, Case, Lin and McLanahan 2000), homicide (Daly and Wilson 1988), and migrants' remittances (Bowles and Posel 2005).

Even for within-household allocations, however, kin altruism is far from an adequate explanation cooperative behaviors. In some studies of food sharing in small scale societies kin-altruism effects appear to be very modest or virtually absent (Gurven, Hill, Kaplan, Hurtado and Lyles 2000a, Kaplan and Hill 1985b, Smith, Bird and Bird 2002). In modern societies important aspects of intergenerational inheritance patterns do not conform to the expectations of the inclusive fitness model in that the (genetically unrelated) spouse typically gets a very substantial bequest, and children typically re-

ceive equal shares irrespective of their age, health status and other correlates of their rreproductive value.

Our study (with Dorrit Posel) of migrants' remittances to their rural families of origin in South Africa (Bowles and Posel 2005) suggests that less than a third of the remittances sent home can be explained by the relatedness of the migrant to the members of the household.

3.3 Reciprocal Altruism

Reciprocal altruism, the second workhorse model for the explanation of cooperation by biologists, concerns repeated interactions among unrelated individuals. Consider a pair of individuals, each of whom in each period will require aid from the other with a certain probability. The behavior whereby each individual provides aid as long as this aid has been reciprocated by the other in the past, is called *reciprocal altruism*. Reciprocal altruism is self-regarding, because each individual's decisions depend only on the net benefit the individual enjoys from the long-term relationship.

Suppose the benefits and costs of the interaction are as above (b and c), and the two available strategies are to offer no assistance (with no benefits or costs) or a contingent cooperation strategy, namely, offer assistance in the first period and in subsequent periods adopt the strategy played by the other in the previous period. Then, as Axelrod and Hamilton (1981) showed, abstracting from time discounting, and assuming that after each round of play the interaction will be terminated with probability t, so that the expected duration of the game is $1/t$ periods, if members of a large population are randomly paired to interact as above, the contingent cooperation strategy will be a best response to itself if $1 - t > c/b$. When compared to Hamilton's rule, this condition makes it clear that the repetition of the interaction (which happens with probability $1 - t$) is analogous to genetic relatedness as a support for cooperative behaviors. If this condition holds in a population composed virtually entirely of conditional cooperators, unconditional defectors will have lower fitness than contingent cooperators, so the contingent cooperation strategy is uninvadable. In this interaction, mutual defect is also an equilibrium. However, reciprocal altruism in the form of contingent cooperation can piggy-back on kin altruism to get started. If the interacting pairs are genetically related, and if the termination probability is sufficiently low, contingent cooperation can invade an all-defect equilibrium, thus accounting for the emergence of contingent cooperation when

initially rare. Once the contingent cooperation strategy becomes prevalent, its persistence does not require relatedness among the interacting pairs.

Biologists have considered reciprocal altruism to be a common form of cooperation among non-kin in the animal world since Trivers' (1971) celebrated article on the subject. Because cooperation based on reciprocal altruism was thought to be ubiquitous in the nonhuman animal world, it was natural for biologists and social scientists to accept it as the standard explanation for human cooperation as well. However, there is little compelling evidence of reciprocal altruism in animal societies.

Peter Hammerstein (2003) writes, for instance, "After three decades of worldwide research on reciprocal altruism and related phenomena, no more than a modest number of animal examples have been identified." A major impediment to self-interested cooperation is that reciprocal altruism requires that animals have a high discount rate (they are extraordinarily impatient), so that future rewards are highly valued in the present. This is virtually never the case (Clements and Stephens 1995, Hammerstein 2003). Behaviors that at first appeared to be reciprocal have frequently, on further study, been better explained as simple mutualism in which the benefit to the actor compensates for the cost of the action irrespective of the action taken by the other (Dugatkin and Mesterton-Gibbons 1996, Milinski 1996, Stephens, McLinn and Stevens 2002).

Consider the well-known study by Wilkinson (1984) of vampire bats, in which a bat that has secured a blood meal will at times share with a non-relative who has failed to secure a meal. It is by no means obvious that reciprocal altruism is involved. Wilkinson showed that if bat A successfully solicits a donation from bat B, there is a greater than chance probability that bat B in the past successfully solicited a donation from bat A. This does not imply, however, that if B refuses A at one time, A will be more likely to refuse B at some future time. For instance, A and B may each, independently, donate to all bats that share some common characteristic with them. Indeed, the vampire bats shared only with roost-mates, who derive from a small number of matrilineal lines, and hence who are biologically related, and not with strange bats. This alone could account for the correlated giving behavior.

Other behaviors that have been interpreted as fitting the reciprocal altruism model do not. Some animals, for instance, exchange several rounds of reciprocal favors over a short period of time. Examples include mutual grooming in the impala (Connor 1995) and mutual egg fertilization in hermaphrodite

fish (Friedman and Hammerstein 1991, Connor 1992). However, in both cases, a more compelling explanation is that when it is one individual's turn to provide the service, it is less costly to do so than simply to free-ride and seek out another partner. Similarly, when small birds find large food sources, they often call out to others, thus promoting the sharing of food. However, the same behavior can be explained by the finder's desire to lower the risk of predation while feeding by attracting additional sentinels and predator targets (Alcock 1993, p. 517ff; Krebs and Davies 1993, p. 120ff). We conclude that while cooperation is common in some species, the basis is most often mutualism or kinship, and that extra-kin reciprocal altruism is rare, sporadic, difficult to maintain except under carefully controlled laboratory conditions, and probably not very important to the life-cycle of nonhuman animals (Stevens and Hauser 2004).

Non-human primates may eventually provide exceptions to this generalization. There is some evidence (surveyed in DeWaal and Davis 2003) that reciprocity plays a role in alliance-formation among macaques and chimpanzees, in the latter among unrelated individuals in high-risk power struggles, both in the wild as well as among captive animals. Capuchin monkeys have been observed experimentally to share more food with a partner whose assistance was needed to acquire the food (de Waal 2000) and chimpanzees to give food to those who groomed them in the past (de Waal 1997). The strongest evidence for reciprocal altruism among non-human primates is from experiments by Hauser, Chen, Chen and Chuang (forthcoming) among unrelated cotton-top tamarin monkeys (*Sauguinus oedipus*). Not only do these animals give more food to a trained conspecific who regularly gives them food, compared to one who does not, they do not reciprocate if the food supplied by the other is observed to have been the byproduct of the other's self-interested actions. They conclude

> …if one tamarin gives another food without obtaining any immediate benefit, then the recipient is more likely to give food in return…[T]amarins altruistically give food to genetically unrelated conspecifics, discriminate between altruistic and selfish actions, and give more food to those who give food back. Tamarins therefore have the psychological capacity for reciprocally mediated altruism.

Among humans the experimental and field evidence for the contribution of repeated interactions to cooperation is overwhelming and well known. An

experiment by Gachter et al. (2004) suggests its importance. Four hundred and fifty six Swiss students played 12560 games (as in section 2.3) in which they were randomly assigned to two roles similar to those of employer and employee in a natural setting. The 'employer' offered a wage, a deduction from his profits, and the ' employee' responded with a level of 'effort,' which was costly to provide. The contrast between two treatments is instructive. In the 'stranger treatment' the pairs were shuffled every period, so that each period was a one-shot interaction for the participants, who were certain they would not encounter any partner more than once. In the 'partner treatment' the two remained paired over ten periods, and this set of ten periods was itself repeated three times. The dominant response for a self-regarding 'employee' in the stranger treatment was to offer one unit of effort. As in the Fehr et al. (1997) experiment reviewed in section 2.3, 'employers' made wage offers far more generous than the minimum required to illicit the one unit of effort. Figure 3.1 gives the effort responses by 'employees' in the two treatments over time.

As expected from the previous chapter, the effort offered even in the stranger treatment is much higher (four times higher) than would have been optimal for a self-regarding 'employee'. More important, the repeated interaction resulted in much higher levels of effort than the stranger treatment, and effort rose over the three sets of treatments and also, except for a dramatic endgame drop-off, within the three sets of play. The fact that repetition contributed to cooperation and that the subjects readily understood the difference between repeated and non-repeated interactions is evident by the sharp reduction during the last two periods of play. The endgame drop-off cannot be due to learning, as high (indeed higher) levels of effort are restored when the second and third set are initiated. The fact that effort did not fall to the level of the stranger treatment suggests that while repetition engaged the self-regarding motives stressed by the reciprocal altruism model, it also tapped social preferences that the stranger treatment did not evoke.

Among humans, therefore, we do not doubt the importance of repeated interactions and other structures that reward cooperators with higher fitness or other payoffs, rendering seemingly selfish acts a form of mutualism. While an important part of the explanation of human cooperation, there are several reasons for doubting the adequacy of this explanation. First, reciprocal altruism fails when a social group is threatened with dissolution, since members who sacrifice now on behalf of group members do not have a high probability of being repaid in the future (see Chapter 6). Second, many human

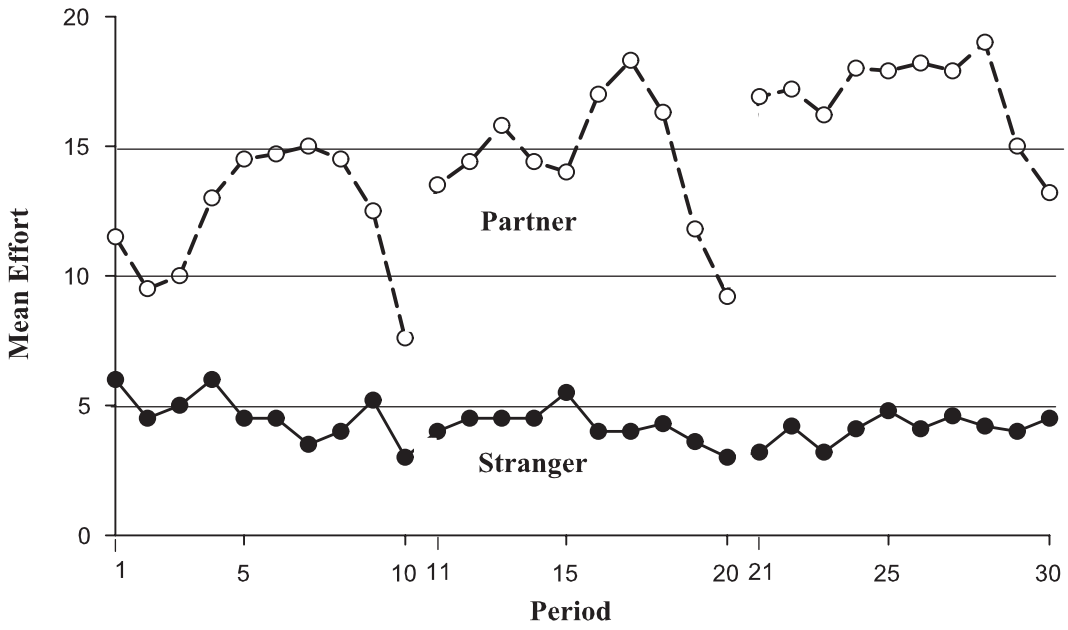

Figure 3.1. Repetition Supports Cooperation in a Gift Exchange (Wage-Effort) Experiment. Based on Gächter, Kessler, and Konigstein (2004)

interactions in the relevant evolutionary context took the form of n-person public goods games—food sharing and other co-insurance, upholding social norms among group members, information sharing, and common defense—rather than dyadic interactions. As we show in section 3.4, it is difficult to sustain cooperation in public goods games by means of the standard tit-for-tat and other reciprocal behaviors (Boyd and Richerson 1988, Joshi 1987) in groups of appreciable size, even if interactions repeated with high probability. Third, the contemporary study of human behavior has documented a large class of social behaviors inexplicable in terms of reciprocal altruism. For instance, there is extensive support for income redistribution in advanced industrial economies, even among those who cannot expect to be net beneficiaries (Fong et al. 2005). Under some circumstances group incentives for large work teams are effective motivators even when the opportunity for reciprocation is absent and the benefits of cooperation are so widely shared

that a self-interested group member would gain from free-riding on the effort of others (Ghemawat 1995, Hansen 1997, Knez and Simester 2001). Finally, as we have seen, laboratory and field experiments show that other-regarding motives are frequently robust causes of cooperative behavior, even in one-shot, anonymous setting.

3.4 Reciprocal Altruism in Large Groups

Among the various criticisms of reciprocal altruism, perhaps the most critical is the inability of this mechanism to foster cooperation in all but the smallest groups. Since reciprocal altruism models are dyadic, it is not even clear how one would extend them to groups with more than two players. In this section we show that the most plausible extension, in which individuals cooperate in a repeated public goods game with $n > 2$ players if a sufficient number of other players cooperated on the previous round, is generally inefficient and evolutionarily unstable.

To this end, we will develop an *agent-based* rather than an analytical model. This is the first of several agent-based models developed in several chapters of this book. Agent-based modeling is important because many dynamic strategic settings are too complicated to admit closed-form analytical solutions. Moreover, the assumptions made to permit explicit analytical solutions to the models are sufficiently unrealistic (e.g., continuous time, infinite numbers of potential individuals) that the agent-based model behaved more like the real-life situations we are trying to model than does a tractable analytical model (Durrett and Levin 1994). The structure and logic of agent-based models is explained in Appendix I to this chapter.

Suppose a large population forms N groups of n members each, and each group plays a public goods game repeatedly d times. We will call this series of rounds an *encounter*. At the end of each encounter, players reassort randomly into groups of size n and carry out another encounter. This continues for P periods, where in our simulations, P varies from 6000 to 25000. By cooperating, a player confers a benefit of b on the other members (i.e., a benefit of $b/(n-1)$ per other member) and a cost c to himself. We assume there are $n+1$ types of players, called m-cooperators, for $m = 0, \ldots, n$. An m-cooperator cooperates in the current round provided at least m other players cooperated in the previous round. On the first round, all players apply the m-criterion as though everyone cooperated in the previous round (i.e., all types cooperate except the n-cooperator, who defects on all

rounds). We call an n-cooperator, who never cooperates, a *defector*, and we call a 0-cooperator an *unconditional cooperator*. Finally we assume that a player who attempts to cooperate will mistakenly defect or be perceived as having defected with probability $\epsilon > 0$, and a player who attempts to defect actually cooperates or is so perceived with the same probability ϵ. We call ϵ the behavioral error rate.

We created a system with 250 groups of size $n = 2, 3$ playing a public goods game repeatedly for ten periods ($d = 10$), in which by cooperating, an agent contributes 2.5 to the other players at a cost of 1 to himself. We initialize the models by letting $m = 0, 1, \ldots, 10$ with equal probability. At the end of each encounter, 5% of agents are replaced by new agents, using a Darwinian fitness criterion according to which the probability of reproduction (respectively, death) is proportional to the agent's payoff relative to others in the population. We assume a mutation rate of 0.5% per period (e.g., if $n = 2$, 250 of the 5000 agents mutate per encounter, or about one mutation per two groups per encounter), where a mutant switches from cooperator to defector or vice-versa with probability 1/2, and if he is a cooperator, he receives a new m assignment, where $m = 0, 1, \ldots, 10$ with equal probability. Also, we assume a behavioral error rate ϵ of 5%. To promote cooperation in the face of errors, we assume that a cooperator who is perceived as defecting (we assume this perception is shared by all group members, so it is public information) plays cooperate unconditionally for the next two rounds, thus allowing cooperation to be restored.

If $n = 2$, the standard dyadic reciprocal altruism case, we find that even if we initialize the population with 50% Defectors, cooperation quickly rises to about the highest level compatible with the given error and mutation rates. This is exhibited in the upper pane of Figure 3.2. Note that the average rate of cooperation is above 80% and the fraction of defectors averages only about 15%. However, intuitions based on the working of the dyadic case do not extend to larger groups. The lower pane in Figure 3.2 shows that when $n = 3$, cooperation quickly deteriorates to a very low level. This shows clearly that reciprocal altruism need not extend to cooperation in groups of size larger then two.

Why is the three person case so different than the standard dyadic case? In the latter, when one defects to punish another defector, the punishment is uniquely targeted on the initial defector. But in groups larger than two one's retaliatory defection in response to another's failure to contribute punishes not only the miscreant, but also all other members of the group whether they

cooperated or not. Thus for groups larger than two, retaliation by defection is no longer targeted on the behavior one would like to deter.

The extent of cooperation shown in Figure 3.2 depends on the ratio of the benefits of cooperation to the costs, b/c, which in those simulations was 2.5. For values of b/c greater than two higher levels of cooperation result for every group size. But as group size rises, the benefit cost ratio sufficient to support a significant level of cooperation rises to implausible levels, as is illustrated in Figure 3.3. In this figure, 'cooperation' is defined as an average of greater than 50% cooperation being maintained in five consecutive runs of the agent-based model with an initial frequency of 25% Cooperators. Other parameters of the model are as before.

3.5 Indirect Reciprocity

Reciprocal altruism is said to occur when frequent repetition induces self-regarding agents to cooperate. In a sizeable group, however, opportunities to cooperate may devolve upon different subsets of individuals at different times. As a result, the probability that the same subgroup will be convened suitably frequently for reciprocal altruism to be effective may be quite low. To address this problem with the reciprocal altruism model, Alexander (1987) proposed a more general cooperative mechanism, which he termed *indirect reciprocity*. In this model, individuals may acquire a *reputation* for being cooperative or selfish through their performance in subgroups. Self-regarding individuals will then cooperate in a newly formed subgroup only if other group members have a sufficiently strong reputation for cooperating. Thus, each group member who invests in a cooperative reputation by cooperating will be rewarded by on average being paired with more cooperative partners.

This could be an important support for the evolution of cooperation. Panchanathan and Boyd (2003) have developed a very elegant model in which an n-person social dilemma is combined with a dyadic indirect reciprocity game in which agents who do not cooperate in the social dilemma are punished by being permanently in bad standing in the indirect reciprocity game. Panchanathan and Boyd shown that for plausible parameters, if the indirect reciprocity game entails high levels of cooperation, this can be used to induce cooperation in the larger game. In addition, the experimentalists Milinski, Semmann and Krambek (2002) show that this mechanism can work with real human subjects.

But, does the indirect reciprocity model support high levels of cooperation under reasonable conditions? As in the case of reciprocal altruism, indirect reciprocity has only been analyzed for dyadic interactions. However, we shall show that indirect reciprocity sustains cooperation in dyadic interaction only under highly favorable conditions concerning the accuracy of the information available to group members concerning the behavior of individuals. In this respect, as a basis for the evolution of cooperation, indirect reciprocity is even less robust than reciprocal altruism.

Suppose that in each period, with some probability an individual becomes "needy" and another individual has the ability to help. We assume that in any period, the status of being needy and potentially helping is uniformly distributed across members of the group. If help is offered, the needy individual receives a benefit of $b > 0$ and the helper incurs a cost c, with $0 < c < b$. Nowak and Sigmund (1998) formalized Alexander's argument by identifying reputation with an *image score*, a form of public information that an individual could increase by helping when called upon to do so. The authors showed that the strategy of helping if the needy individual has a high image score is stable against invasion by defectors, and weakly stable against invasion by unconditional cooperators once defectors are eliminated from the population. However, Leimar and Hammerstein (2001), showed that the strategy of helping when one's own image score is low, independent of the image score of the needy, can invade and destroy Nowak and Sigmund's cooperative equilibrium, and Panchanathan and Boyd (2003) showed that if small performance errors are introduced into Nowak and Sigmund's original model, universal defect becomes the only stable equilibrium of the system.

Both Leimar and Hammerstein (2001) and Panchanathan and Boyd (2003) identify the problem with image-scoring: there is no incentive for a self-regarding individual to care about the status of the needy when deciding whether or not to help. They suggested an alternative that had been proposed by Sugden (1986), called the *good standing* model of indirect reciprocity. According to this model, every group member has a 'standing' which is public information within the group, and is either 'good' or 'bad.' All individuals start out in good standing. An individual moves from good to bad standing by refusing to help a needy person who is in good standing, and moves from bad standing to good standing by helping a needy individual.[2]

[2]Sugden discusses an alternative, in which an individual moves from bad standing to good standing by helping needy individual only if the latter is in good standing. This

Panchanathan and Boyd show that the "discriminating reciprocator" (r) strategy of helping the needy who are in good standing when the helper is in good standing and always helping the needy when in bad standing is evolutionarily stable in a population where the other strategies are universal defect (d) and universal cooperate (c). However, universal defect is also evolutionarily stable, so the capacity of a group to maintain the cooperative equilibrium using indirect reciprocity depends on the explicit dynamics that will select which of the two equilibria (universal defect or cooperation supported by indirect reciprocity) will obtain. As the authors stress, this depends in turn on the quality of the information concerning cooperation and defection in dyadic interactions.

Following Panchanathan and Boyd (2003), let us assume each individual needs help once in each period and is asked to help once in each period, and that with probability q an individual knows the standing of a needy person he is asked to help, and with probability $1 - q$ he has no information concerning the needy persons status. Suppose in this case a discriminating reciprocator in good standing helps unless he knows the needy individual is in bad standing. Using this model, in Appendix III to this chapter, we find that the information requirements for indirect reciprocity being an evolutionary stable strategy are quite stringent. We show that the minimum feasible q is always greater than 1/2, and increases with α, where $\alpha > 0$ is the probability that an individual accidentally fails (or is perceived to fail) to help. In other words, $q > 1/2$ is a necessary condition for an evolutionarily stable indirect reciprocity equilibrium.

This requirement is extremely strong, and is not likely to be met in most real-world conditions. This is because the dyadic interaction of helper and needy is usually private to the interacting parties, and hence is unlikely to be observed by more than a small number of non-participants. We note that this condition for the stability of a reciprocator equilibrium would be even more demanding if we bowed to realism by permitting private, perceptual errors in the standing signal. However, we should not be surprised at this results, for there are no known examples of indirect reciprocity among non-human society.

It may be objected that behavior consistent with the indirect reciprocity mechanisms identified by Alexander are common in human society. We agree, this could be due to other-regarding preferences, and not to the self-

definition leads to only minor differences in the behavior of the model, so we will not discuss this alternative here.

regarding preferences assumed by Nowak and Sigmund, and Panchanathan and Boyd. Engelmann and Fischbacher (2004) analyzed an indirect reciprocity setting experimentally under two treatments. In the public treatment information about one's helping behavior towards others on previous rounds was made available to other subjects in a public image score. In the private treatment no information concerning ones helping behavior was made available to other subjects. Though less than in the public treatment, a significant amount of helping behavior was observed even in the private treatment. This shows that not all of the helping behavior observed in the experiment is explained by the expectation of future reward for one's generous behavior. Given the presence of subjects with other-regarding preferences, one cannot determine how much of the helping behavior can be attributed to the self-regarding motives stressed in the indirect reciprocity model.

3.6 Altruism as a Signal of Quality

A final form of mutualistic cooperation is that which results when a cooperative act is a difficult-to-fake signal of some characteristic of the actor that is otherwise difficult to observe. In this case cooperative behaviors may be favored in evolution because they enhance the individual's opportunities for mating and coalition building. This would be the case, for example, if sharing valuable information or incurring dangers in defense of the group were taken by others as an honest signal of the individual's otherwise unobservable traits as a mate or political ally. In this case self-regarding individuals might engage in group-beneficial activities in anticipation of reproductive, political, or other benefits. Models of this process were developed by Spence (1973), initially applied to educational attainment as a signal, and Zahavi (1977) initially applied to helping behaviors among Arabian babblers. Much of the literature on costly signaling and human evolution explains such behaviors as good hunters contributing their prey to others (Smith and Bliege Bird 2000, Sosis 2000). The same reasoning applies to cooperative behaviors.

Cooperative behaviors would thus result in advantageous alliances for those signaling in this manner, and the resulting enhanced fitness or material success would then account for the proliferation of the cooperative behaviors constituting the signal. We have modeled this process as a multi-player public goods game that involves no repeated or assortative interactions, so that non-cooperation would be the dominant strategy if there were no signaling benefits (Gintis, Smith and Bowles 2001). We show that honest

signaling of underlying quality by providing a public good to the rest of the group can be evolutionarily stable and can proliferate in a population in which it is initially rare, provided that certain plausible conditions hold. Behaviors conforming to what we call strong reciprocity thus could have evolved in this way.

Our signaling equilibrium alone, however, does not require that the signal confer benefits on other group members. Anti-social behaviors could perform the same function—beating up one's neighbor can demonstrate physical prowess just as much as behaving bravely in defense of the group. If signaling is to be an explanation of group beneficial behavior, the logic of the model must be complemented by a demonstration that group beneficial signaling is will be favored over antisocial signaling. We supply this by noting that the level of public benefit provided may be positively correlated with the individual benefit the signaler provides to those who respond to the signal. For instance, it may be that the signaler who defends the group is more likely to confer a benefit (say, protection) on his partner or allies than the signaler who beats up his neighbor. Group-beneficial signals such as sharing one's prey may attract larger audiences than anti-social signals. Finally, multilevel selection among competing groups would favor groups at beneficial signaling equilibria over those either at non-signaling equilibria or those at anti-social signaling equilibria.

As this last reason suggests, the effects of signaling and multilevel selection on cooperation may be synergistic rather than simply additive. Multilevel selection provides a reason why signaling may be pro-social, while signaling theory provides a reason why group beneficial behaviors may be evolutionarily stable in a within-group dynamic, thus contributing to between group variance in behavior and thereby enhancing the force of multilevel selection.

Costly signaling has been proposed as an explanation for certain types of food-sharing in human societies, such as providing game that is large and/or difficult to harvest, or large quantities of food for consumption at ritual feasts (Boone 1998, Gurven, Allen-Arave, Hill and Hurtado 2000b, Hawkes, O'Connell and Blurton Jones 2001, Smith and Bliege Bird 2000, Sosis 2000). In most such cases, as in the chimpanzee case, there is not sufficient information available to judge if key conditions for costly signaling, such as quality-dependent signaling, are present. In the case studied by Bliege Bird and colleagues (Bliege Bird, Smith and Bird 2001), such data are now available, and agree with costly signaling predictions. Young men are more likely to establish reputations for foraging ability by providing large game (marine

turtles) for feasts attended by upwards of 200 people than by other means, such as conspicuous consumption or minimizing foraging time to supply domestic needs, which would only be observed by immediate neighbors, and even for them would be less conspicuous than feast contributions.

3.7 Positive Assortation Permits Cooperation to Flourish

The models discussed thus far assume that groups are formed by random draws from the population. We have seen that with this assumption it is extremely difficult for a large group facing a social dilemma to sustain a high level of cooperation using plausible informational and payoff assumptions. In this section, we show that by relaxing the assumption of random association, a high level of cooperation among agents can be sustained under an expanded range of conditions.

We assume the conditions of the model presented in Section 3.4, except that we now add that each agent is also a *k-associator*, where $k = 0, 1, \ldots, d$. When a group has completed an encounter of d_1 rounds, suppose the agent who had the highest rate of cooperation in the encounter is a k-associator. This agent then forms a new group including all the agents in his group who cooperated at least kd_1/d times, plus a number of new agents drawn randomly from the population of agents who were not invited to remain with their group members, sufficient to restore the size of each group to n. In effect, those who cooperate at the rate of k/d or higher remain together for at least one additional encounter.

We simulated this model, as before, with a benefit-cost ratio of 2.5 to 1, and letting $k = 0, 1, \ldots, d$ with equal probability both initially and when mutations occur. We find that even if with group size $n = 30$ and all groups initialized with all defectors, cooperation quickly rises to about the highest level compatible with the given error and mutation rates. This is exhibited in Figure 3.4. Note that the average rate of cooperation is above 80% and the fraction of defectors averages only about 9%.

We conclude that cooperation can be maintained by positive assortation even in large groups under the appropriate conditions. The critical assumption is that cooperators can positively assort at no cost. If groups occupy areas with scarce resources that are controlled by the groups, by the definition of 'scarce,' excluded group members will not be assured of quickly finding a

new group, and hence will resist exclusion. Even in the absence of scarcity, a group with many cooperators will be a desirable location for a defector, who will then also resist exclusion. If excluding defectors is costly, however, a second-order free rider problem arises. An individual who cooperates but does not incur the costs of excluding defectors will have higher fitness than a cooperator who incurs such costs. Handling this free rider problem is quite as serious as handling the first-order free rider problem posed by the public goods game, and hence positive assortation of cooperators is no solution at all. In Chapter 5 we show that expelling uncooperative group members may evolve even when it is costly to those carrying out this form of punishment, as long as sufficiently many group members are strong reciprocators.

3.8 Conclusion

Biologists, and others inspired by biological reasoning, have provided a number of plausible models of the process by which cooperation could have evolved. Among these are kin-based altruism, reciprocal altruism, indirect reciprocity, and positive assortation. While important in explaining many aspects of human cooperation, these theories do not provide an adequate account of the phenomenon. Kin altruism is important in relationships among relatives. But other than possibly being the template for more generalized forms of altruism towards unrelated individuals, it does not explain one of the most distinctive aspects of human cooperation, namely that taking place among unrelated individuals. Reciprocal altruism provides a plausible account of some forms of cooperation among dyads, but not in larger groups and not when future interactions are unlikely. Indirect reciprocity facilitates cooperative behavior, but were preferences entirely self-regarding it would not support cooperative outcomes except under quite implausible assumptions concerning what individuals know about the behavior of members of their group when interacting with others. Positive assortation arising from the deliberate choices of people not to associate with those who have not behaved cooperatively toward them in the past can promote cooperation among large numbers of non-kin under plausible information assumptions. But, in the model showing this, it was assumed quite implausibly that one could refuse to associate with another member of one's group at no cost to oneself.

Nonetheless, two important truths about the underpinnings of cooperation have been made clear by these models. The first, most clearly shown by

the model of reciprocal altruism, is that sustaining cooperation requires the punishment of those who exploit the cooperation of others. The second, evident from models of kin selection and positive assortation among non-kin, is that cooperation may be supported by population structures that result in non-random pairing of individuals, such that likes interact with likes more often than would occur by chance. The essential contributions of punishment of defectors and positive assortation due to the group structure of populations are the subject of Chapters 5, 6, and 7. We will see that the evolution of other-regarding preferences is crucial to the explanations offered in these chapters, and their centrality in our account of human cooperation is the main distinction between our view and those we have just summarized.

Before showing how the punishment of defectors and positive assortation can support the distinctly human forms of cooperation among large numbers of non-kin, we must explore refinements in the modeling of mutualistic cooperation developed in the last two decades by economists. We will see that these contributions go considerably beyond the simple models described here. But like the biologically inspired models, they do not provide an explanation of why large numbers of self-regarding individuals with noisy and sometimes private information about one another's activities might nonetheless cooperate in social dilemmas. We will conclude that cooperation on a human scale requires that at least some individuals be other-regarding.

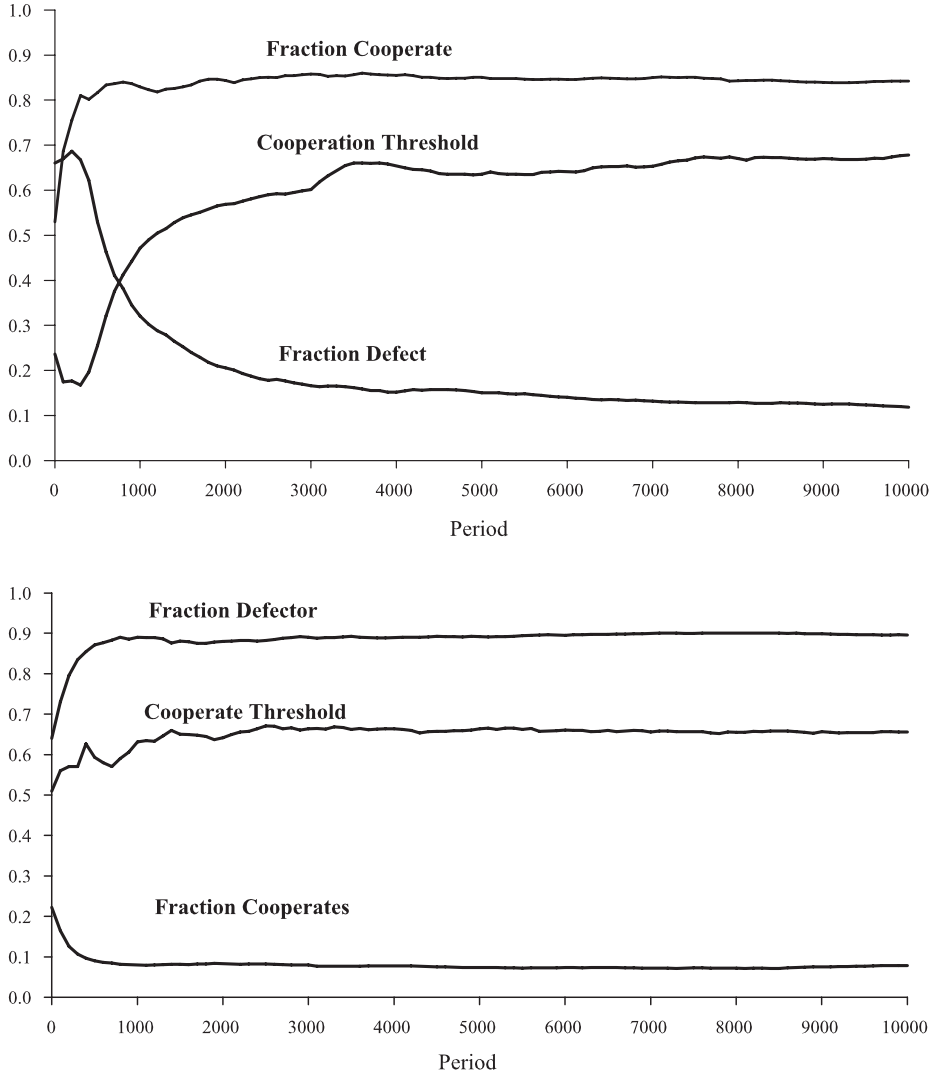

Figure 3.2. Cooperation Ten-round Public Goods Game using Reciprocal Altruism Strategies. We used a mutation rate of 0.5% per period, an error rate ϵ of 5%, and a benefit/cost ratio of 2.5 to 1. The upper pane assumes $n = 2$, the traditional reciprocal altruism case, and the lower pane assumes $n = 3$. 'Fraction Defect' is the fraction of moves that are defects, 'Fraction Cooperate' is the fraction of moves that are cooperate, and 'Cooperate Threshold' is the average fraction of players that must cooperate on one round in order that a cooperator cooperate on the subsequent round.

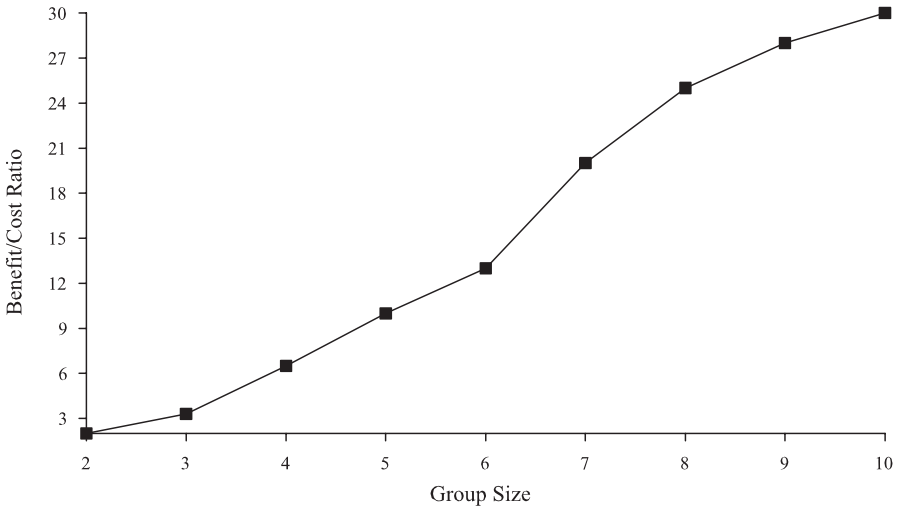

Figure 3.3. Group size and Minimum Benefit/Cost Ratio Permitting Cooperation with Reciprocal Altruism. 'Cooperation' here means that with an initial frequency of 25% Cooperators, an average of greater than 50% cooperation is maintained in five consecutive runs of the agent-based model.

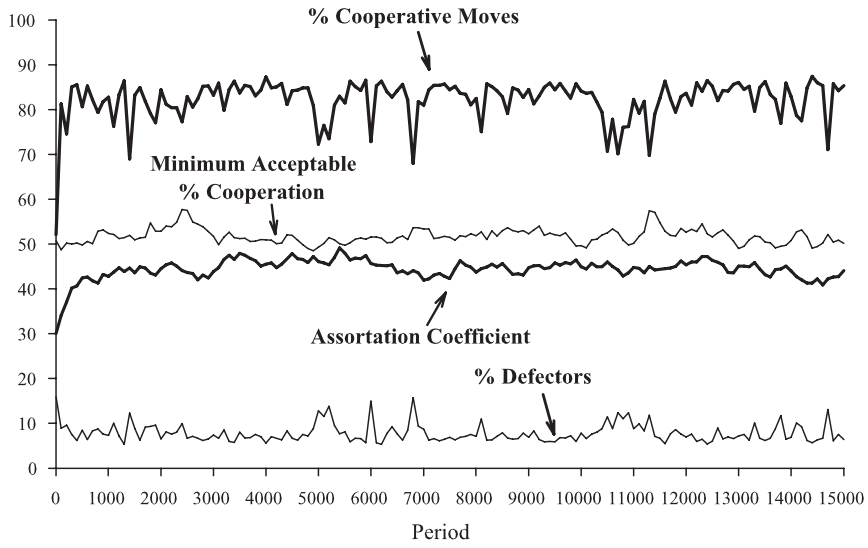

Figure 3.4. Cooperation in Thirty-player, Ten-round Public Goods Game where Cooperators can Costlessly Positively Assort. 'Minimum Acceptable % Cooperators' is the average level of cooperation in the previous round below which a cooperator will defect in the current round. The 'Assortation Coefficient' is the minimum frequency of cooperation and individual must exhibit in the previous encounter to be accepted into a cooperator's group in the next encounter.

4

Cooperative *Homo economicus*

> The first principle of economics is that every agent is actuated only by self-interest.
>
> F. Y. Edgeworth, *Mathematical Psychics (1881), p. 104*

4.1 Introduction

A major goal of economic theory since the second half of the Twentieth century has been to show the plausibility of wide-scale cooperation among self-interested individuals. The first thrust in this endeavor involved the rigorous specification of Walras' general equilibrium model. Despite the success of this endeavor (Arrow and Debreu 1954, Debreu 1959, Arrow and Hahn 1971), the essential assumption that contracts could completely specify all relevant aspects of all exchanges and could be enforced at zero cost to the exchanging parties is widely recognized as not applicable to many important aspects of economic life. Indeed, such economic institutions as firms and state agencies depend strongly on incentive mechanisms involving strategic interaction in addition to explicit contracts, a fact long ago recognized and studied by sociologists (Blau 1964, Gintis 1976, Granovetter 1985).

The second major thrust eschewed complete contracting and involved the development of sophisticated repeated game-theoretic models of strategic interaction. These models refine and extend Trivers' insight that retaliation against defectors by withdrawal of cooperation may enforce cooperation among self-regarding individuals. A great virtue of these models is that in recognizing the ubiquity of incomplete or unenforceable contracts they describe the real world of interactions among most animals, including humans (Bowles and Hammerstein 2003).

These models are an important contribution because they confirm the speculative assertions of Shubik (1959), Trivers (1971), Taylor (1976), and Axelrod and Hamilton (1981). The most complete statement of this line of

thinking, carefully applied towards understanding the broad historical and anthropological sweep of human experience, is the work of Ken Binmore (Binmore 1993, Binmore 1998), culminating in his *Natural Justice* (2005).. This work offers an evolutionary approach to morality, in which moral rules form a cultural system that grew historically with the emergence of *Homo sapiens* and is evolutionarily robust and stable. Successful social rules, Binmore stresses, must be *efficient* in the sense that societies that use these rules prosper at the expense of those that do not, and they must be *fair* or they will not be accepted by those whom such rules govern. For Binmore, a society's moral rules are instructions for behavior in conformity with one of the myriad of Nash equilibria of a repeated n-player social interaction. Because the interactions are repeated, the self-regarding individuals who comprise the social order will conform to the moral rules of their society as a type of self-fulfilling prophecy (if all other individuals play their part in this Nash equilibrium, an individual has no incentive to deviate from playing his part as well). Binmore's solution is one of a broad class of models developed in the economics literature to explain cooperation among self-regarding agents as a results of repeated interactions. In this chapter, we will show that none of these models is successful.

In this chapter, we identify three shortcomings of such repeated-game models. First, for cooperation to be sustained, individuals must have implausible cognitive capacities and levels of patience. Second, for groups of any appreciable size, the amount and quality of information required for these models to support high levels of cooperation will typically not be available. Third, the cooperative equilibria described by the models are unlikely to be hit upon by any plausible evolutionary process, and if attained would not persist for long; i.e., they are unstable in plausible dynamical frameworks.

Because the quality of information will play an important part in our discussion, let us clarify some terms. Information is said to be *imperfect* if it is noisy; that is, if the signal concerning some individual's action is not perfectly correlated with the action. Information is *perfect* if for all individuals, the signal is perfectly correlated with the action. Information is said to be *private* if individuals do not receive the same information, that is, if the signal concerning some individual's action received by two individuals may be different. Information is *public* if all relevant signals received by all pairs of individuals are perfectly correlated. If the forager who should be hunting with the others is stretched out under a tree, information about this is private if not all of the band members have this information; it is imperfect

if the signal they receive is that he is shirking, when in fact he may be resting after a particularly fatiguing and unsuccessful pursuit of a dangerous prey.

We begin this chapter with a list of *desiderata* for what would constitute an adequate explanation of cooperation. We then take up the main contributions to this literature beginning with trigger strategies and then considering targeted punishment before turning to the 'industry standard' in this field, the Fudenberg, Levine and Maskin model. We then introduce private information before concluding. This chapter is somewhat more demanding mathematically than the rest of the book. The next section and the conclusion will provide an overview of its main arguments.

4.2 Requirements for a Model of Cooperation

A *public goods game* is an n-person prisoner's dilemma in which, by "cooperating," each individual A adds more to the payoff of the other members than A's cost of cooperating, but A's share of the total gains he creates is less that his cost of cooperating. By "defecting," the individual incurs no personal cost and produces no benefit for the group.

Consider a group of size n, where each member can work or shirk in each time period $t = 1, 2, \ldots$. The cost of working is $c > 0$ and the benefit is $b > c$, shared equally among the other group members. A self-interested member will shirk in a one-shot encounter because it costs c to cooperate and the benefits accrue only to others. However, suppose the encounter is repeated in each period, and all members have discount factor δ, with $0 < \delta < 1$. Suppose also that with probability $\epsilon > 0$, a member who chooses to work will fail to produce the benefit b, and will appear to the other members of the group to have shirked.

Now consider the expected benefits and costs of working in this situation. If $\hat{k}$ others work, a member expects to receive $\hat{k} \times b(1-\epsilon)/(n-1) = b\nu(1-\epsilon)$ from their effort, where $\nu = \hat{k}/(n-1)$. The present value of working is then $v_c = b\nu(1-\epsilon) - c + \delta v_c$, the first two terms on the right hand side representing the immediate net benefit of cooperation, and the final term the discounted benefit of future returns. Solving for v_c, we get

$$v_c = \frac{b\nu(1-\epsilon) - c}{1-\delta}. \tag{4.1}$$

An individual who does not work, by the same reasoning, earns

$$b\nu(1-\epsilon)/(1-\delta) > v_c. \tag{4.2}$$

Clearly, *each self-regarding individual will defect, and total payoff will be zero*. Repeated game models deal with this problem by imposing costs on the defector sufficient to offset the gain from shirking. We evaluate such models on the basis of the following criteria.

First, cooperation must be sustained without third-party contract enforcement or other unexplained social institutions. This so-called *incentive compatibility requirement* reflects not only the universal condition of humanity prior to some 10,000 years ago, but also the condition of most work groups in a modern economy that must solve its incentive and coordination problems without recourse to courts or police, except perhaps under extreme circumstances. In particular, this requirement implies that any payoff redistributions dictated by the rules for cooperation, as well as any punishments meted out, must be effected by individuals who maximize utility subject only to the constraints and incentives imposed by the behavior of other group members. In particular, if individuals are to be punished (by fine, social exclusion, or physical harm), the punishers must have an incentive to carry out the punishment.

Second, an acceptable model must show that cooperation is a *plausible evolutionary outcome*. Random fluctuations in costs and payoffs, as well as errors of commission (e.g., attempting, but failing to cooperate), signal transmission (e.g., appearing to defect, while in fact having cooperated, and *vice-versa*), and the periodic introduction of novel strategies (by mutation or other processes) must not disrupt cooperation or entail excessive efficiency losses. Thus, the model must be *evolutionarily stable* with respect to standard dynamical forces (i.e., more successful strategies tend to grow at the expense of less successful strategies), under a variety of plausible conditions, in the sense of being impervious to invasions by mutant individuals or by groups that deploy competing forms of social organization. Finally, the proposed model must be capable of forming and growing in an indifferent or hostile environment in which cooperation is initially rare.

Third, the organizational forms and incentive mechanisms deployed in the model must reflect the types of strategic interaction and incentives *actually observed in human groups*. In particular, the model should work well with group sizes on the order of at least ten to twenty, and the incentives to punish defectors should reflect those deployed in real-world public goods game settings.

Fourth, the information required by group members for cooperation to be sustained must be empirically plausible. Signals concerning the behavior of

group members should be *imperfect* in the sense that the signal indicating the cooperation of particular individual may be incorrectly received with moderate probability (say, 5% or 10% for any individual in any period), and *private* in the sense that the signals indicating the cooperation of particular individual may be uncorrelated across other individuals in the group, conditional on the actual performance of the signaler. This requirement follows from the fact that in observed public goods game settings, group members generally observe only a subset of other group members in any one period.

Fifth, cooperation should be sustained under empirically *plausible rates of discounting the future*. The probability of death in any given period places an upper bound strictly less than one on individual discount factors. Even restricting attention to the years of low mortality between childhood and senescence, demographic estimates from modern day Hadza and Aché foragers suggest that our distant ancestors probably experienced average mortality of 2% per year in the prime of life, giving an upper bound for the discount factor of 0.98 (Kaplan et al. 2000). Moreover, empirical estimates of discount factors in industrial societies are well below this upper bound, including estimates based on consumer purchases of durable goods (Hausman 1979) as well as laboratory experiments Green, Myerson, Lichtman, Rosen and Fry (1996). These data also show that there is a statistical distribution of discount factors among members. For various reasons, the probability of future interactions will differ across group members, and there is individual variation in discount rates across individuals. Moreover, within the same individual, the discount rate will vary across time and personal circumstance. An acceptable model must therefore function effectively in the face of plausible statistical distribution of discount factors. Because the distribution of discount factors in a group has positive variance, the common practice in the literature of repeated game theory of taking the limit as the discount factor goes to unity is inappropriate.

We will show that to date, all models of cooperation based on self-interested individuals violate one or more of these conditions, and hence fail to solve the problem of cooperation among unrelated individuals.

4.3 Trigger Strategies

If people are self-interested, according to the celebrated *folk theorem* of repeated game theory, cooperation can be sustained if all individuals use the following *trigger strategy*: cooperate as long as all other team members

cooperate. We illustrate this with the public goods game with imperfect information as just introduced. Whenever a member defects (or appears to defect), defect for a sufficient number of periods, say k, so as to render the defector worse off than if he had cooperated. Cooperation is a best response for all players provided they are sufficiently patient, and if a defection is observed, defecting for the required number of rounds is also a best response. Suppose the cooperate/defect decision of each member is an imperfect public signal, so that with probability $\epsilon > 0$ an intended cooperation will appear to be a defection. The value of cooperating when all other members cooperate is now given by the recursion

$$v_c = b(1-\epsilon) - c + \delta(1-\epsilon)^n v_c + (1-(1-\epsilon)^n)\delta^k v_c.$$

The first two terms capture the immediate net payoff, the third term is the discounted future returns to cooperation assuming no erroneous signals of defection were transmitted, multiplied by the probability no erroneous signals were transmitted, and the final term is the discounted future return to cooperation in the case that at least one erroneous signal was detected, so that k periods of non-cooperation occur prior to the resumption of cooperation, multiplied by the probability that an erroneous signal was transmitted. Solving this equation, we get

$$v_c = \frac{b(1-\epsilon) - c}{1 - \delta^k - \delta(1-\delta^{k-1})(1-\epsilon)^n}. \tag{4.3}$$

The present value of defecting is $v_d = b(1-\epsilon) + v_c\delta^k$. For cooperation to be a subgame perfect Nash equilibrium it must be that$v_c \geq v_d$, or

$$\frac{c}{b} \leq \frac{\delta(1-\delta^{k-1})(1-\epsilon)^{n+1}}{1-\delta^k}, \tag{4.4}$$

so if

$$c > b(1-\epsilon)^{n+1}, \tag{4.5}$$

the cooperative equilibrium cannot be sustained for any k, no matter how patient the group members (i.e., no matter how close δ is to unity). Indeed, (4.4) shows that *an error rate of ϵ in a group of size n is equivalent to reducing the discount factor by a factor of* $(1-\epsilon)^{n+1}$. Thus, no matter how small the probability ϵ, if the group is sufficiently large, cooperation cannot be sustained.

It is also easy to check that, assuming k is chosen to minimize the cost of sustaining cooperation, so that (4.4) is an equality, the average per-period payoff to members is

$$v_c = \frac{b(1-\epsilon)}{1-\delta^{k+1}}. \tag{4.6}$$

The right hand side of this expression is approximately $b(1-\epsilon)$ for plausible values of δ and for even moderate values of k.

For example, suppose $n = 15$, $b = 1.5$, $c = 1.0$, and $\delta = 0.95$. Then the highest error rate that can be sustained is $\epsilon \approx 2.2\%$. At $\epsilon = 2.0\%$, we must set $k = 19$ for a cooperative equilibrium, and the punishment stage, lasting 20 periods long, occurs with 20% probability after each period. In other words, on average, for every five periods of cooperation there will be 20 periods of noncooperation.

The inefficiency in this model stems from the fact that, given the self-interested nature of the actors, the only way to punish a defector is to cease cooperation completely for a sufficiently long period of time. As we showed through agent-based models in Section 3.4, when the group is large, this is clearly an inefficient form of punishment—guilty and innocent suffer equally and with a plausible error rate, most of the groups' time is spent idling. We will see that targeted punishment by self-regarding individuals does not solve the problem.

4.4 Directed Punishment

Suppose we institute the direct punishment of defectors, where a cooperative strategy includes inflicting an amount $p > 0$ of punishment on any individual who defected, at a cost $c_p > 0$ to the punisher. Giving the benefit of the doubt to the model, we assume defection is detected with certainty by each group member, and $p > c > c_p$, so punishment is relatively cheap to inflict, and relatively costly to sustain. In this case, by defecting, an individual incurs punishment cost $np + c_p$ (assuming for simplicity that the punishee is obliged to punish himself), so the condition for cooperating is $np + c_p \geq c$. We assume these conditions hold (the self-punishing simplification is odd, but it does not materially affect the results).

If punishment is costly, and assuming a symmetric equilibrium in which all individuals use the same punishment probability, self-interested individuals will punish in Nash equilibrium only if $nq^*p + c_p = c$, where q^* is the probability of punishing a defector. Certainly such a $q^* \in (0, 1)$ exists,

and leads to a cooperative mixed strategy Nash equilibrium with no need to enforce the decision to punish. This equilibrium, however, has extremely poor stability properties, since an individual who deviates by punishing with probability slightly less than q^* leads to the immediate collapse of cooperation. Moreover, such deviations would surely occur because if all members adopt q^*, then for each, the expected payoffs to punishing all of the time, or none of the time, will be identical to the payoffs expected by playing the mixed strategy q^*. This instability can be avoided by choosing q to be strictly greater than q^*, but then the chosen strategies do not form a Nash equilibrium. We should note that this instability is not due to the assumption of symmetric strategies. More complex rules for determining who punishes under what conditions lead to the same instability of the Nash equilibrium.

An alternative is to create incentives for individuals to punish defectors. Suppose members agree that any individual who is detected not punishing a defector is himself subject to punishment by the other players. Suppose with probability ϵ an individual who intends to punish fails to do so, or is perceived publicly by the other members to have failed. For simplicity, we choose ϵ to be the same as the error rate of cooperation. If all individuals cooperate and punish, the rate of observed defection will be ϵn, so the mean number of punishment of defector events per period will be ϵn^2. But, of course, $\epsilon^2 n^2$ of these events will erroneously be viewed as non-punishing, so we must have $\epsilon^2 n^3$ punishing of non-punisher events (let us call this *second order punishment*). Similarly, we must have $\epsilon^3 n^4$ third-order punishment to enforce second-order punishment. Assuming we have punishment on all levels, the total amount of punishment per individual per period will be

$$\epsilon n(1 + \epsilon n + \epsilon^2 n^2 + \ldots) = \frac{\epsilon n}{1 - \epsilon n},$$

provided $\epsilon < 1/n$. If the reverse inequality holds, this mechanism cannot work because each order of punishment involves greater numbers than the previous. Assuming $\epsilon < 1/n$, the expected payoff to a cooperator under conditions of complete cooperation (assuming one engages in one's own punishment) is given by the recursion equation

$$v_c = b(1 - \epsilon) - c - \epsilon n \frac{p + c_p}{1 - \epsilon n} + \delta v_c,$$

so

$$v_c = \frac{(b(1 - \epsilon) - c)(1 - \epsilon n) - \epsilon n(p + c_p)}{1 - \delta}, \tag{4.7}$$

which becomes negative when ϵ is sufficiently close to $1/n$. Thus, cooperation will not be sustainable for large n (say, > 20) unless error rates are implausibly low.

Another possible punishment strategy is to punish second-order nonpunishers using a trigger strategy, for which the infinite recursion problem does not occur. While this would involve a higher rate of employing the trigger strategy, if the cost of second-order punishment is lower than that of first-order punishment, the length of the mutual defection period will be shorter. In this case, the mean number of punishment events per period is ϵn^2. The probability that no individual will be punished for nonpunishing is thus $(1-\epsilon)^{\epsilon n^2}$. The recursion equation for complete cooperation is then

$$v_c = b(1-\epsilon) - c - \epsilon n(p + c_p) + \delta(1-\epsilon)^{n^2\epsilon} v_c + \left(1 - (1-\epsilon)^{n^2\epsilon}\right) \delta^{k+1} v_c,$$

where k is the number of periods of universal defection needed to ensure that defecting in second-order punishing is unprofitable. Solving, we get

$$v_c = \frac{b(1-\epsilon) - c - n(c_p + p)\epsilon}{1 - \delta^{k+1} - \delta(1-\delta^k)(1-\epsilon)^{n^2\epsilon}}, \tag{4.8}$$

which is decreasing in ϵ and n, and increasing in δ. The gain from not punishing for one period and then returning to cooperation is $b(1-\epsilon) - c - n\epsilon p + \delta^{k+1} v_c$. The gain from punishing over nonpunishing is thus positive when

$$b(1-\epsilon) - c - np\epsilon \geq \frac{1-\delta^{k+1}}{1-\delta^k} \frac{n\epsilon c_p}{\delta(1-\epsilon)^{n^2\epsilon}}. \tag{4.9}$$

Since the first fraction on the right hand side of this equation is greater than unity, a necessary condition for a cooperative equilibrium is then

$$b(1-\epsilon) - c > \frac{n\epsilon(p + c_p)}{\delta(1-\epsilon)^{n^2\epsilon}}. \tag{4.10}$$

This will clearly be violated for sufficiently large n, for any give $\epsilon > 0$ and $\delta < 1$. However, we can now have $\epsilon > 1/n$, as the next example shows.

Suppose the parameters of the model are as in the previous numerical example except $c_p = c/n$ and $p = 2c/n$ (i.e., it is not very costly to punish nonpunishers, but it is costly to be punished for nonpunishing). Then it is easy to calculate that for $\epsilon = 1/n \approx 0.067$, we have $k = 3$. The probability of completing a cooperative period without the need for punishing

nonpunishers is $(1-\epsilon)^{n^2\epsilon} \approx 0.355$, but the payoff rate is less than 2.5% of the payoff possible if compliance were costlessly enforced. With $\epsilon = 1/2n \approx 0.034$, the probability of completing a cooperative period without the need for punishing nonpunishers is $(1-\epsilon)^{n^2\epsilon} \approx 0.755$, and the payoff rate is about 2/3 of the payoff possible if compliance were costlessly enforced. Clearly, then for very low error rates this compliance mechanism can be effective, but with error rates above 4%, its effectiveness deteriorates sharply.

4.5 The Classical Game-Theoretic Model

Our investigation thus far has found that when substantial numbers of self-regarding individuals have the option of providing public goods, and where information involving contributions and punishment of defecting members is public but imperfect, the benefits of cooperation that are realized through repeated interactions are small or even nonexistent for plausible group sizes and error rates. Even the existence of a cooperative equilibrium requires implausible error rates, discount rates, and group sizes. Moreover, the strategies commonly introduced to support cooperation in the standard repeated game models—triggers and second order punishment—are implausible on empirical grounds. An important model due to Fudenberg, Levine and Maskin (1994) shows that levels of cooperation at or near Pareto-efficient levels may be sustained, seemingly overcoming the deficiencies we have identified above.

In this section we will outline the Fudenberg, Levine and Maskin analysis (hereafter FLM) as applied to the public goods model laid out above, and show why it does not solve the problem maintaining cooperation among self-interested individuals.

This model is highly attractive in that is shows that for any number n of sufficiently patient individuals (δ sufficiently close to unity), and any error rate $\epsilon > 0$, the repeated game can support close to a Pareto optimum, using trigger strategies alone (i.e., by reacting to perceived defections by some pattern of defection of other members, plus cooperation of the guilty parties, of sufficient duration to render purposeful defection unprofitable). This is in sharp contrast to the highly error-sensitive existence condition (4.5) and payoff schedule (4.6).

The FLM Nash equilibrium involves many technical details that we need not develop here—an analytical description of the model is presented in the Appendix I to this chapter, in Section 12.4. The main idea, however, is

very simple. If everyone cooperates in every period, average per-member payoff per period is $b^* = b(1-\epsilon) - c$, so the total payoff per member is $b^*/(1-\delta)$. If one individual is detected shirking, all agree to work a little less hard in every succeeding period, such that the total discounted value from shirking is c. This ensures that cooperating is a best response. One way to achieve this is to have all individuals work with probability $p_\epsilon = 1 - c(1-\delta)/b^*$ in each period, then after the detection, the total loss of payoff is $b^*(1-p_\epsilon)/(1-\delta) = c$. Thus, there is no gain to shirking. Moreover, when δ is close to unity, p_ϵ is close to unity, so the per-period payoff $b^* p_\epsilon$ is very close to b^*, so cooperation is close to efficient.

However, a member will in fact work with probability p_ϵ after the infraction only if this is a best response. Thus, all members must agree to a new, lower, probability of working that renders shirking unprofitable. Clearly, this leads to an infinite decent, with a new agreement in effect after every detected defection. But, FLM show that for a given $\epsilon > 0$, there is always a δ sufficiently close to unity that the average payoff per period per member is as close to b^* as desired. The resulting Nash equilibrium is called a *sequential equilibrium* because at every possible decision point, even at ones that cannot occur in the Nash equilibrium (specifically, decision points where one of the members has defected), each member plays a best response to the play of the other members.

Though reliant on empirically implausible trigger strategies, the model thus seems to have overcome many of the other deficiencies of models that sustain cooperation among self-regarding individuals by game repetition. But this is not the case: its dynamic properties make the equilibrium virtually irrelevant from an evolutionary standpoint, the information requirement for sustaining an equilibrium even if attained are extremely stringent, and the discount factors required to sustain cooperation are implausible. We consider these three problems in turn.

First, the fact that a pattern of activity is a sequential Nash equilibrium says nothing whatever about its dynamic properties. Sequential Nash equilibria are impervious to single individuals playing alternative strategies, whether on or off the game path. Dynamic stability, by contrast, requires that the model be capable of recovering from small, simultaneous, perturbations in the strategies of *all individuals*. This is certainly not the case for the FLM model, simply because there is an open set of sequential equilibria, meaning that in every neighborhood of such equilibrium, there are uncountable other such equilibria. This implies that no such equilibrium can be locally asymp-

totically stable, since no such equilibrium can have a basin of attraction of positive radius.

Second, the FLM model has extremely strong, indeed quite implausible, informational requirements.[1] Perhaps the most obvious problem is that there are *many* sequential equilibria, not just one, since the necessary reduction in cooperation dictated by the detection of a defection can be shared in many different ways among the members, and each member has an incentive to grab as large a share of the total as possible. Similarly, if there is some shock to the system, there is no mechanism for the restoration of any particular one of the many sequential equilibria that ensure a high level of cooperation. In short, there is no dynamic that picks out any particular solution, and no dynamic that ensures that any solution can respond effectively to stochastic disturbances. If we add the fact that there may be some uncertainty concerning the value of c for each member of the group, or that without a public randomizing device (e.g., a big roulette wheel that all use, and is publicly observable, to decide when to defect on purpose) it may be difficult to check that a member actually uses the probability p_ϵ appropriate to the sequential equilibrium.

Third, while it is of great interest to know that a certain Nash equilibrium exists for sufficiently patient individuals, we cannot expect a real social group to have extremely patient members. Note that because condition 3.11 must hold for all members of the group, it is the least patient that determines if cooperation will be sustained.

4.6 Cooperation with Private Information

In the models discussed to this point, each member of the group receives the same, perhaps imperfect, signal of the behavior of each other member.

[1]Consider, for instance, the well-known sufficient conditions for Nash equilibrium of Aumann and Brandenburger (1995), which can be stated as follows: *Let $\mathcal{G}$ be a game with n players and let σ be an n-dimensional strategy profile. Suppose the players have a common prior, which assigns positive probability to it being mutually known that $\mathcal{G}$ is the game being played, mutually known that all players are rational, and commonly known that σ is the strategy profile being played. Then σ is a Nash equilibrium.* We say $\sigma = \{\sigma_1, \ldots, \sigma_n\}$ is *mutually known* if each individual knows σ, and σ is *commonly known* if each individual knows σ, each knows that the others know σ, and so on. A player is *rational* if he maximizes his payoff subject to beliefs.While these conditions are not necessary, they have been shown to be strict (i.e., with a violation of any one, a non-Nash equilibrium counterexample can be found). There can be no substantive necessary conditions for Nash equilibrium, since a random strategy profile can, by accident, represent a Nash equilibrium.

In most empirically relevant cases, however, different group members will receive different, and perhaps conflicting, signals concerning other members.

One of the most important developments in economic theory in the past decade is a rigorous analysis of cooperation with private signaling. It may be thought that simply pooling information would transform private into public information. This is the case, however, only if we assume that individuals report their private information truthfully. However, without the proper incentives for truth-telling, we cannot plausibly make this assumption. The situation is especially serious when the private information is also imperfect. For, in this case, observing a defection, one does not know whether others observed it, and hence if C observes the "failure" of B to punish the defection of C, this could occur when A did not defect, or when A did defect but B did not witness the defection, or when A defected, B did witness the defection, but C mistakenly failed to witness B's punishment of A.

The obvious next step is consider more general ways to use private information efficiently. This is in fact the tack taken in recent years by several economists (see Kandori, 2002 for an overview). Important contributions to this research agenda include Sekiguchi (1997), who was the first to propose a Nash equilibrium that approximately sustains the cooperative payoff in the two-person prisoner's dilemma, assuming that private monitoring is nearly perfect. Following this, contributions by Piccione (2002), Ely and Välimäki (2002), and Bhaskar and Obara (2002), and Matsushima (2000) considerably deepened the approach.

The technical problems involved in developing an equilibrium with a high level of cooperation assuming private information and self-interested individuals are extreme when the private signals are imperfect. If punishing an observed breech of cooperative norms is costly, and if team members generally do not know which members observed which breeches, costly first-order punishment will not occur because those who see the defection know that they will not be punished for failing to punish. Therefore, in order to occur, first-order punishment must not be costly. The various ways of achieving this result involve the use of mixed strategy sequential equilibria, so these models are vulnerable to the critique that the mechanisms involved are not seen empirically and are not evolutionarily stable.

We will shall sketch two quite distinct models in this tradition. These are the most powerful of their respective types, yet each has irremedial weakness as an explanation of cooperation in public goods games. The first, analyzed by Bhaskar and Obara (2002), is a private signal version of the model in

Section 4.3. We revise our earlier model by assuming that when an individual moves in each period, he sends signals to each of the other $n-1$ individuals. These signals are independently distributed, given the move, but each is in error with probability $\epsilon > 0$. We will consider only symmetric equilibria, in which all individuals employ the same strategy, and we assume that after a defection, a player defects forever (this is called the *grim trigger* strategy).

The first complication of the private signal assumption is that no sequential equilibrium can support full cooperation in any period. To see this, consider the first period. If each player uses the full cooperation strategy, then if a player receives a defection signal from another player, with probability one this represents a bad signal rather than an intentional defection. Thus, with very high probability, no other member received a defection signal. Therefore no player will react to a defect signal by defecting, and hence the always defect strategy will have a higher payoff that the always cooperate strategy. To deal with this problem, we must have all players defect with positive probability in the first period. A similar analysis applies to all future periods.

Now, in any Nash equilibrium, the payoff to any two pure strategies that are used with positive probability by a player must have equal payoffs against the equilibrium strategies of the other players. Therefore, the probability of defecting must be chosen so *each player is indifferent between cooperating and defecting on each round*. The analysis of the implications of this fact by Bhaskar and Obara is subtle and ingenious. They prove that under plausible conditions, for any n there is an error rate ϵ^* sufficiently small that, when errors are of frequency ϵ^* or less, there is a sequential equilibrium in which the discounted payoff to individuals is approximately that of the fully cooperative solution.

There are two problems with this solution. First, empirical relevance suggests that the error rate will be given by social and technical conditions, and will be a function of group size: $\epsilon = f(n)$. Thus, it is not reasonable to choose n first, then freely choose ϵ. Indeed, we might expect ϵ to increase with group size. This will occur, for instance, if each member has a fixed frequency of interacting with other members, so as group size increases, the probability of detecting a defection of any particular member decreases. As we have seen in Section 4.3, when we hold ϵ constant, there is a finite limit to the group size that can be supported, and using plausible parameter values, this limit is likely to be quite small. Moreover, as we have seen, even when a Nash equilibrium exists, it is likely to be quite inefficient.

The second problem is that, as we have just explained, there is no reason to believe that a sequential Nash equilibrium be evolutionarily stable. To illustrate this problem, we have constructed an agent-based model of the Bhaskar-Obara model in the Delphi programming language, as implemented by Borland Delphi 6.0 (for an introduction to agent-based models, see Appendix II to this chapter). The stage game is as above, with $b = 3$ and $c = 1$. Individuals are randomly assigned to groups of size n in each of 100,000 periods. In each period, each group plays the stage game repeatedly, the game terminating with probability 0.05 at the end of each round, analogous to a discount factor of 0.95. The program begins by creating 210 individuals, each endowed at the time of creation with two parameters. The first, *DefectRound*, indicates at which round the individual will voluntarily defect. If this is very large, the individual never defects. Since we wish to assess the stability of equilibrium rather than whether it is accessible from an arbitrary initial distribution, the program initially assigns 80% of individuals with $DefectRound = 100$, which effectively means they never defect. The other 20% of individuals are randomly assigned *DefectRound* values between 1 and 10. The second parameter is *Tolerance*, which indicates how many defections an individual who voluntarily cooperates must see before beginning to defect. All individuals are assigned $Tolerance = 1$, so they defect at the first defection signal they receive (this is the equilibrium value for the Bhaskar-Obara model).

In each round, for each group, each member sends a signal indicating whether he cooperated or defected, with error rate ϵ, to every other group members. On the basis of this signal, all individuals then update their willingness to cooperate in the next round. As soon as the round hits or exceeds an individual's *DefectRound*, or he accumulates more than *Tolerance* defect signals, the individual defects from that point on with this particular group.

At the end of every 100 periods, the program implements a reproduction phase, using the relative fitness of the individuals, as measured by their accumulated score over the 100 periods, and replacing 5% of poorly performing individuals by copies of better performing individuals. We implement this by a simple imitation process that has the same dynamic properties as the replicator dynamic (Taylor and Jonker 1978) (see Section 3.5; for a derivation of the replicator dynamic, see Appendix II of Chapter 3). For each replacement, we randomly choose two individuals, and the individual with the higher score is copied into the individual with the lower score.

At the completion of each reproduction phase, the program implements a mutation phase, in which each individual's parameters are increased or decreased by one unit (except if so doing would lead to negative values) with probability 0.001.

As might be expected, when we set $n = 2$, the dynamic process exhibits a high level of efficiency (about 90% of full cooperation), as well as a high level of tolerance (individuals defect after about seven defect signals, on average) even with the quite high error rate of $\epsilon = 10\%$, after 100,000 rounds.

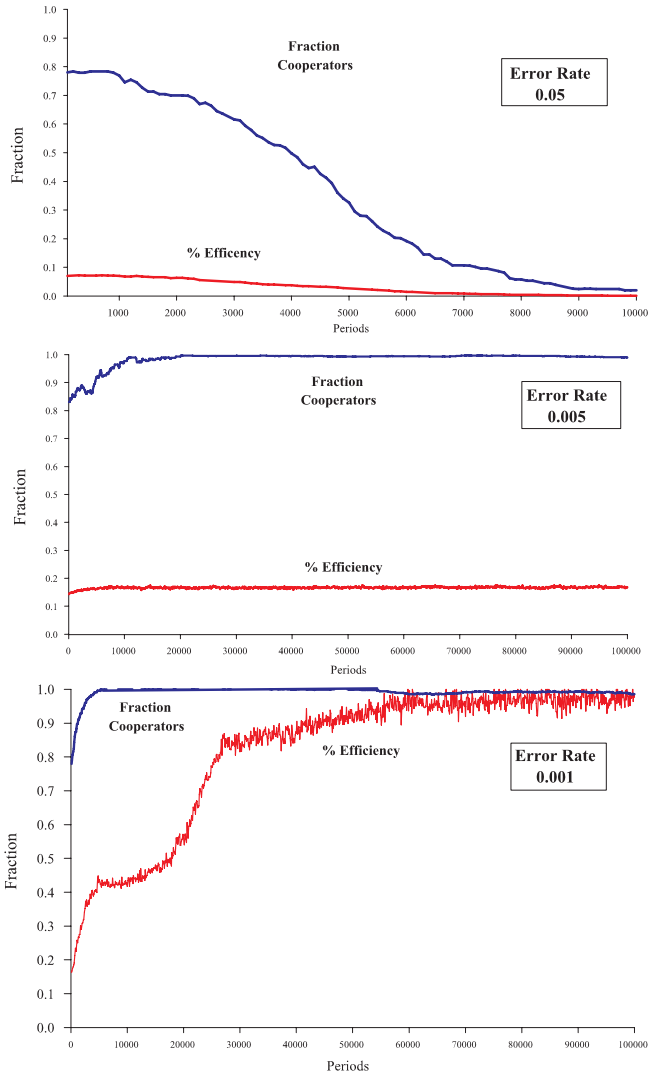

Figure 4.1. Agent-Based Model of Bhaskar-Obara model with group size $n = 10$, with model parameters are as described in the text. The upper pane shows that for the empirically plausible error rate of 5%, there is no cooperation and very low efficiency. Even for the implausibly low error rate of 0.5%, illustrated in the middle pane, there is very low efficiency despite a high prevalence of cooperators. Only when the error rate is reduced to 0.1%, as depicted in the lower pane, does efficiency attain a high level.

When we raise group size to $n = 10$, however the picture is quite different. The upper pane in Figure 4.1 illustrates the case with error rate $\epsilon = 5\%$. Note that even with this relatively small group size, the level of cooperation falls to very low levels. Lowering the error rate to $\epsilon = 0.5\%$, as in the middle pane in Figure 4.1, we see that the level of cooperation becomes high, but the efficiency of cooperation (the ratio of actual payoffs to the maximum attainable if individuals always cooperated) is only about 17%. This is because cooperation is signaled as defection between some pair of individuals with probability close to 40%. Only when we set the error level to $\epsilon = 0.1\%$, as in the lower pane of Figure 4.1, do we achieve a high level of efficiency, the probability of an individual receiving a defection signal when in fact all are cooperating now falling below 10%. Since this low error level also allows a high level of tolerance, defections become quite rare. However, a 0.1% error rate is implausibly low.

Ely and Välimäki (2002) have developed a quite different approach to the problem, following the lead of Piccione (2002), who showed how to achieve coordination in a repeated game with private information without the need for complex belief updating of the sort required by Bhaskar and Obara (2002). They achieve this by constructing a sequential equilibrium in which at every stage, each player is indifferent between cooperating and defecting no matter what his fellow members do. Such an individual is thus willing to follow an arbitrary mixed strategy in each period, and the authors show that there exists such a strategy for each player that ensures close to perfect cooperation, provided individuals are sufficiently patient and the errors are small.

The weakness of this approach in explaining real-world cooperation, which can be stated and supported without explicitly presenting the Ely-Välimäki model, it is one shared by the mixed strategy Nash equilibria of many games. In a general one-shot game, if a player's mixed strategy is part of a Nash equilibrium, then the payoffs to all the pure strategies used with positive probability must be equal. Hence no player has an incentive to calculate and use the mixed strategy at all, since he does equally well by simply choosing among the pure strategies occurring in the support of the mixed strategy in question. If there are costs to computing and randomizing, however small, choosing the most convenient pure strategy will be strictly preferred to computing and playing the mixed strategy.

The intuition behind this result is straightforward. The Ely-Välimäki construction has the explicit goal of making individuals indifferent to the

actual moves of other members, thus admitting a particular constellation of mixed strategies that are specified precisely to implement an efficient, cooperative equilibrium. But, if payoffs are perturbed, players are no longer indifferent and do not have the proper incentives to implement the near-efficient solution.

4.7 Conclusion

The game-theoretic models we have surveyed here have illuminated the conditions under which self-regarding individuals might cooperate. To sustain cooperation, it has been necessary to attribute to the actors extraordinary cognitive capacities and future oriented time horizons, virtually banishing impatience and behavioral or perceptual errors. Yet even if equipped with these capacities, it is unlikely that groups of any size would ever discover the cooperative equilibria that the models have identified, and almost certain, if they were to hit on one, that it would unravel in short order. While there is no question but these models have strongly advanced our understanding of cooperation, we have shown that the cooperative equilibria they describe are exceedingly unlikely to have emerged in any realistic evolutionary setting, and had they done so would quickly have been abandoned. Thus these models are irrelevant in a dynamic setting. The relevance of these models is further compromised by a tendency to look only at the limiting characteristics of solutions as discount rates approximate unity and error rates approach zero. Finally, no attempt has ever been made to show that such models apply to the major forms of cooperation in firms, states, neighborhoods, and elsewhere. There is considerable doubt that they could so apply, since plausible discount factors and error rates will tend to unravel cooperation.

By contrast, the models based on social preferences to which we now turn support cooperation as an evolutionarily plausible outcome, do not depend on discount factors near unity, and permit relatively high error rates in private signals.

Our critical assessment in this chapter concerns a class of models, the Folk theorem-inspired reasoning we have reviewed, not the entirely correct underlying idea that among individuals with self-regarding preferences, repeated interactions may promote cooperation. Like the models of kin-altruism and costly signaling reviewed in the previous chapter, the reciprocal altruism allowed by repeated interactions provides part of the explanation of human cooperation.

We showed in Section 3.7 that if cooperators can selectively associate with other cooperators at no cost to themselves then in repeated interactions in groups of substantial size, cooperation can be sustained. In that model, individuals made use only of the private information they had obtained in previous interactions with the members of their own group. The model thus meets our requirements for an adequate explanation of cooperation. But the assumption that individuals can exclude others from a group at no cost to themselves is quite implausible in most cases. We therefore turn in our next chapter to a setting in which exclusion of noncooperators is costly.