Prosocial Emotions*

Samuel Bowles and Herbert Gintis

June 21, 2002

Abstract

Adherence to social norms is underwritten not only by the cognitively mediated pursuit of self-interest, but also by emotions. Shame, guilt, pride, regret, joy and other visceral reactions play a central role in sustaining cooperative relations, including successful transactions in the absence of complete contracting. Prosocial emotions function like the basic emotion, "pain," in providing guides for action that bypass the explicit cognitive optimizing process that lies at the core of the standard behavioral model in economics. We consider a public goods game where agents maximize a utility function that captures five distinct motives: personal material payoffs, one's valuation of the payoffs to others, which depend both on one's altruism and one's degree of reciprocity, and one's sense of guilt or shame in response to one's own and others' actions. We present empirical evidence suggesting that such emotions play a role in the public goods game, and we develop an analytical model and an agent-based simulation showing that reciprocity, shame, and guilt increase the level of cooperation in the group. Finally, we provide an explanation of the long term evolutionary success of prosocial emotions in terms of both the individual and group-level benefits they confer.

*We dedicate this paper to Kenneth Arrow, both as a scientist and a person, for whom we have the deepest admiration, and from whom we have drawn the deepest inspiration. Presented at the workshop, Economy as a Complex Evolving System, III, in honor of Kenneth Arrow, Santa Fe Institute, November 16-18, 2001. Thanks to Kenneth Arrow, John Geanakoplos, Charles Manski, Giorgio Topa, Peyton Young and other workshop participants for helpful comments, to George Cowan for the van Gogh quote, to the John D. and Catherine T. MacArthur Foundation for financial support, and the Santa Fe Institute for a stimulating research environment.

Let's not forget that the little emotions are the great captains of our lives and we obey them without realizing it.

Vincent Van Gogh in a letter to his brother Theo

The heart has reasons that Reason knows nothing about.

Blaise Pascal, Pensées *(1670)*

How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it. Our imagination therefore attaches the idea of shame to all violations of faith.

Adam Smith, The Theory of Moral Sentiments *(1759)*

1 Introduction

Social interactions in modern economies are typically quasi-contractual. Some aspects of what is being transacted are regulated by complete and readily-enforceable contracts, while others are not. Transactions concerning credit, employment, information, and other goods and services where quality is difficult to monitor provide examples of quasi-contractual exchanges. Where contracting is absent or incomplete the optimality properties of decentralized market allocations no longer hold. But where the invisible hand fails, the handshake may succeed. Kenneth Arrow, who we honor with this essay and this volume, wrote (1971):22

> *In the absence of trust…opportunities for mutually beneficial cooperation would have to be foregone…norms of social behavior, including ethical and moral codes [may be]…reactions of society to compensate for market failures.*

As in many other areas, Arrow's insight long predates the recent recognition of the economic importance of norms. Surprisingly little progress has been made in the intervening years in understanding how norms affect behavior and why some norms that impose costs on their adherents, such as forgoing opportunities to lie, cheat, and steal even when the prospect of discovery is vanishingly small, might have been successful by the test of either genetic or cultural evolution. This lack of progress, we think, may be traced to two shortcomings of the way behavioral scientists have addressed the problem. The first is the common representation of seemingly unselfish acts as reflecting the far-sighted pursuit of self interest. The second is the neglect of emotions as important influences on behavior.

An explanation of the adherence to social norms with wide acceptance in biology (Trivers 1971), evolutionary psychology (Cosmides and Tooby 1992), political science (Taylor 1976, Axelrod and Hamilton 1981), and economics (Fudenberg and

Maskin 1986) is that individually costly behaviors that confer benefits on others are sustained by the repeated nature of interactions that allow for punishment of norm violators. We have explained elsewhere why we believe these explanations to be insufficient. In brief, they fail to explain compelling evidence of adherence to norms in both experimental and real world situations that are clearly nonrepeated. Moreover, in interactions among more than a few individuals, it is very difficult to sustain high levels of adherence to social norms if errors in play or in the perceptions of others' play occur (Boyd and Richerson 1988, Bowles and Gintis 2001).

A second reason for our limited success in understanding social norms is the remarkable neglect of emotions in the study of behavior. It may seem odd that an approach once said to be based on the "calculus of pleasure and pain" would pay so little attention to feelings. But in the standard economic model actions are taken to bring about valued consequences. The process by which the individual arrives at the action is cognitive, not affective. Visceral reactions such as joy, shame, fear, and disgust thus play no role in the process of decision making, however much their anticipation may influence the evaluation of the consequences of an action. The neglect of the behavioral consequences of emotions is not limited to economics, but extends to psychology and neuroscience as well, where cognitive aspects of behavior is a major line of research, while the causes of emotions receive far more attention than their behavioral consequences.[1]

The interpretation we would like to advance here is that adherence to social norms is underwritten by emotions, not only by the expectation of future reciprocation. The experience of shame, guilt, pride, regret, joy and other visceral reactions plays a central role in sustaining cooperative relations, including successful transactions in the absence of complete contracting. An example will illustrate our view and its potential relevance to economic policy making.

Parents are sometimes late in picking up their children at day care centers. In Haifa, at six randomly chosen centers a fine was imposed for lateness while in a control group of centers no fine was imposed (Gneezy and Rustichini 2000). The expectation was that punctuality would improve at the first group of centers. But parents responded to the fine by even greater tardiness. The fraction picking up their children late more than doubled. Even more striking was the fact that when after 16 weeks the fine was revoked, their enhanced tardiness persisted, showing no tendency to return to the *status quo ante*. Over the entire 20 weeks of the experiment, there were no changes in the degree of lateness at the day care centers in the control group. The authors of the study, Uri Gneezy and Aldo Rustichini, reason that the fine was a contextual cue, unintentionally providing information about the appropriate

[1]This situation is being rectified. In psychology, see Zajonc (1980) and Damasio (1994), and in economics see Loewenstein (1996), Laibson (1996), and Bosman and van Winden (2001).

behavior. The effect was to convert lateness from the violation of an obligation which might have occasioned the feeling of guilt, to a choice with a price that many were willing to pay. They titled their study "A Fine is a Price" and concluded that imposing a fine labeled the interaction as a market-like situation, one in which parents were more than willing to buy lateness. Revoking the fine did not restore the initial framing, but rather just lowered the price of lateness to zero.

The fact that monetary incentives for punctuality instead induced even greater tardiness is both counter to the predictions of the standard behavioral model in economics and suggests an alternative approach in which social norms and the activation of emotions when they are violated play a central role in behavior. We define a behavior as *prosocial* if its exercise increases the average payoff to members of the group. One of the most important emotions contributing to prosocial behavior is *shame*, the feeling of discomfort at having done something wrong not only by one's own norms but also in the eyes of those whose opinions matter to you.[2]

Prosocial emotions function like the basic emotion, "pain," in providing guides for action that bypass the explicit cognitive optimizing process that lies at the core of the standard behavioral model in economics. Antonio Damasio (1994):173 calls these "somatic markers." A somatic marker is a bodily response that "forces attention on the negative outcome to which a given action may lead and functions as an automated alarm signal which says: Beware of danger ahead if you choose the option that leads to this outcome.…the automated signal protects you against future losses." Emotions thus contribute to the decision-making process, not simply by clouding reason, but in beneficial ways as well. Damasio continues: "suffering puts us on notice.…it increases the probability that individuals will heed pain signals and act to avert their source or correct their consequences." (p. 264)

To explore the role of guilt and shame in inducing prosocial behaviors we will consider a particular interaction having the structure of a public goods game. We assume individuals maximize a utility function that captures five distinct motives: one's individual material payoffs, how much one values the payoffs to others, which depend both on ones' altruism and one's degree of reciprocity, and one's sense of guilt or shame in response to one's own and others' actions. To this end, we will amend and extend a utility function derived from the work of Geanakoplos, Pearce and Stacchetti (1989), Falk and Fischbacher (1998), Levine (1998), and Sethi and Somanathan (2001).

The shame term in the utility function captures the idea that individuals may experience discomfort based on their beliefs about the extent to which it is socially

[2]Shame differs from guilt in that while both involve the violation of a norm, the former but not the latter is necessarily induced by others knowing about the violation and making their displeasure known to the violator.

acceptable to take self-interested actions at the expense of others. The sense of shame is not exogenously given, but rather is influenced by how others respond to one's actions. Thus an individual taking an action that generates a personal material payoff while inflicting costs on others may provoke punishment by fellow group members resulting in a reduction in payoffs of the miscreant. But in addition to the payoff reduction, he also may experience a level of shame that depends, in addition to the action he took, the extent to which other group members expressed their disapproval by inflicting punishment upon him.

In the public good setting, contributing too little to the public account may evoke shame if one feels that has appropriated "too much" to oneself. Because shame is socially induced, being punished when one has contributed little triggers the feeling of having taken too much In this case, the effect of punishment on behavior may not operate by changing the incentives facing the individual, that is by making it clear that his payoffs will be reduced by the expected punishments in future rounds. Rather it evokes a different evaluation by the individual of the act of taking too much, namely, shame. This is the view expressed by Jon Elster (1998):67 "material sanctions themselves are best understood as vehicles of the emotion of contempt, which is the direct trigger of shame." Thus, self-interested actions, *per se*, may induce guilt, but not shame. If one contributes little and is not punished, one comes to consider these actions as unshameful. If, by contrast, one is punished when one has appropriated very little, the emotional reaction may be spite towards the members of one's group.

The interpretation of behavior advanced here may be contrasted with a related and complementary modification of the canonical behavioral model in economics, namely, the assumption of bounded rationality (Simon 1982). In our interpretation, agents may be deviating from the predictions of the standard model not because they are incapable of doing the cognitive operations required by the model but because they do not feel like doing (and acting on) these calculations. Indeed their feelings may cause them to act in ways inconsistent with the standard model even when they have flawlessly done the required calculations.

In Section 2, we present experimental evidence consistent with the view that punishment not only reduces material payoffs but also recruits emotions of shame towards the modification of behavior in prosocial directions. In Section 3, we model of the process by which an emotion such a shame may affect behavior in a simple three-person public goods game. In Section 4, we generalize to an n-person public goods game. In Section 5, we ask how behaviorally important emotions such as shame might have evolved. We conclude with some implications for economic theory and policy.

2 The Moral Response to Punishment: Experimental Evidence

Strong reciprocity is the predisposition to cooperate with others and punish non-cooperators, even when this behavior cannot be justified in terms of self-interest, however broadly conceived. An extensive body of evidence suggests that a considerable fraction of the population, in many different societies, and under many different social conditions, including complete anonymity, are strong reciprocators. We here review laboratory evidence concerning the public goods game. For additional evidence, including the results of dictator, ultimatum, common pool resource and trust games, see Güth and Tietz (1990), Roth (1995), and Camerer and Thaler (1995).

The public goods game consists of n subjects under conditions of strict anonymity. Each subject is given w 'points,' redeemable at the end of the experimental session for real money. Each subject then places some number of points in a 'common account,' and keeps the rest. The experimenter then gives each subject a fraction $q \in (1/n, 1)$ times the total amount in the common account. Contributing is thus an altruistic act, because it increases the average payoff to the group ($q > 1/n$) at the expense of the individual ($q < 1$).

Contributing nothing to the common account is a dominant strategy in the public goods game if subjects are self-interested. Public goods experiments, however, show that only a fraction of subjects conform to the self-interested model. Rather, subjects begin by contributing on average about half of their endowment to the common account.

If the game is continued over several rounds, however, contributions tend to fall. In a meta-study of twelve public goods experiments Fehr and Schmidt (1999) found that in the early rounds, average and median contribution levels ranged from 40% to 60% of the endowment, in the final period (usually round ten) 73% of all individuals ($N = 1042$) contributed nothing, and many of the remaining players contributed close to zero. The explanation of the decay of cooperation offered by subjects when debriefed after the experiment is that cooperative subjects became angry at others who contributed less than themselves, and retaliated against free-riding low contributors in the only way available to them—by lowering their own contributions (Andreoni 1995). Experimental evidence supports this interpretation. When subjects are allowed to punish noncontributors, they do so at a cost to themselves (Dawes, Orbell and Van de Kragt 1986;Sato 1987; Yamagishi 1988a,b,1992; Ostrom, Walker, and Gardner, 1992).

Fehr and Gächter (2000), for instance, set up a ten round public goods game with $n = 4$ and costly punishment, employing three different methods of assigning members to groups. Under the *Personal* treatment, the four subjects remained in the same group for all ten periods. Under the *Stranger* treatment, the subjects

were randomly reassigned after each round. Finally, under the *Perfect Stranger* treatment the subjects were randomly reassigned and assured that they would never meet another subject more than once (in this case, the number of rounds had to be reduced from ten to six to accommodate the size of the subject pool). Subjects earned an average of about $35 for an experimental session.

Fehr and Gächter (2000) performed their experiment for ten rounds with punishment and ten rounds without. Their results are illustrated in Figure 1. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the Partner game, despite strict anonymity, cooperation increases almost to full cooperation, even on the final round. When punishment is not permitted, however, the same subjects experience the deterioration of cooperation found in previous public goods games.

The contrast between the Partner effect and the two Stranger effects is worth noting. In the latter case punishment prevented the deterioration of cooperation, whereas in the former case punishment led to an increase in participation over time, until near full cooperation was achieved. This result suggest that subjects are motivated by the personal desire to punish free riders (the Stranger treatment), but are even more strongly motivated when they there is an identifiable group, to which they belong, whose cooperative effort is impaired by free riding (the Partner treatment). The prosociality of strong reciprocity is thus more strongly manifested, the more coherent and permanent the group in question.

The frequency with which subjects paid to punish other group members raises serious doubts about the adequacy of the standard behavioral model, for in the perfect stranger treatment (or in the final periods of other treatments) the dominant strategy is to contribute nothing and to refrain from punishing. Indeed, strategically, punishment is identical to the contribution to the public good. Both are forms of altruism—a benefit conferred on others at a cost to oneself. The fact that subjects avidly punish low contributors, and display considerable negative affect when asked why they do so, suggests that they are responding emotionally—specifically, they are acting on feelings of anger.

We focus in this paper on the response of the punishees, which appears no less prompted by emotions. Unlike punishing behavior, which cannot be motivated by payoff gains, a positive response to the experience of being punished could be explained by the desire to avoid further reductions in payoffs due to being punished in subsequent rounds. But as we will see, in many experimental setting, this motivation explains only part of the response. We will first present one of our own experiments conducted with Jeffrey Carpenter (Bowles, Carptenter and Gintis 2001) and then comment on the results of two remarkable experiments by others.

By implementing the Stranger Treatment, in which subjects are randomly reassigned to a new group at the beginning of each round of play, we deliberately created

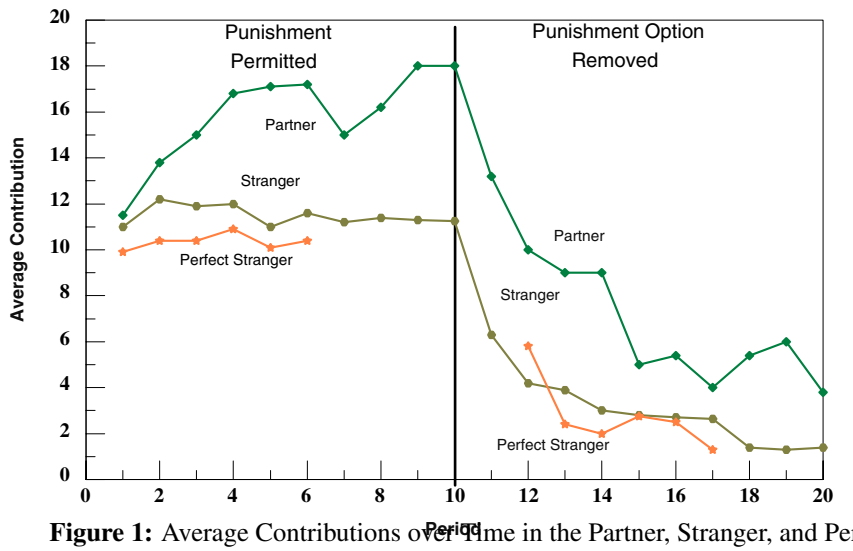

Figure 1: Average Contributions over Time in the Partner, Stranger, and Perfect Stranger Treatments when the Punishment Condition is Played First (adapted from Fehr and Gächter, 2000).

an experimental environment in which cooperation would be difficult to sustain.[3] We also make punishing shirkers quite costly to punishers: the cost of inflicting a penalty of two experimental "points" is one point for the punisher. Suppose there are n players. Each player receives w points at the beginning of each round, and player i contributes a_i to the public good. These contributions are revealed to the other players, who then can punish by purchasing as much punishment as they want at a cost of one point per sanction. Let μ_{ij} be the expenditure on sanctions assigned by player i to player j (we assume $\mu_{ii} = 0$). Then the payoff to player i is given by

$$\pi_i = w - a_i + q \sum_{j=1}^{n} a_j - \sum_{j=1}^{n} \mu_{ij} - 2 \sum_{j=1}^{n} \mu_{ji}. \tag{1}$$

Note that the first two terms $(w - a_i)$ represent the amount i keeps for himself, the third term is the amount he receives from the common pool, the fourth term is the amount he spends on punishing others, and the final term is the amount he is punished by others.

[3]The more common Partners Treatment, in which groups remain together throughout the experiment, tends to foster more cooperation than the Stranger Treatment (Croson 1996).

To study the effect of group size and the degree of harm caused by shirking, we used two group sizes (four and eight) and two values of q (0.3 and 0.75), allowing us to compare across our treatment variables to look for similarities in behavior based on the punishment that shirkers inflict on other group members. Our underlying behavioral assumptions concerning reciprocity imply that an agent's punishment of another agent would vary both with the other agent's shirking rate and the harm caused by a unit of shirking, the latter depending on the size of the group and the marginal per-person return on contribution to the public account. There are two ways to measure the harm done by a shirking group member. The first, which we term the *private cost of shirking* is the reduction in each agent's payoffs associated with an act of shirking by individual i, or $q(w - a_i)$. By contrast, z_i, the social cost of shirking by member i takes account of the costs borne by every group member other than the shirker, or $(n - 1)q(w - a_i)$.

We conducted twelve sessions, three per treatment, with 172 participants. The number of participants, and therefore groups, per treatment vary due to no-shows. All subjects were recruited by email from the general student population and none had ever participated in a public goods experiment before. Each subject was given a five dollar show-up fee upon arrival and then was seated at a partially isolated computer terminal so that decisions were made in privacy. Each session took approximately 45 minutes from sign-in to payments and subjects earned $20.58 on average, including the show-up fee.

Each session lasted ten periods. In each period (a) subjects were randomly reassigned to a group, given an endowment of $w = 25$ points, and allowed to contribute, anonymously, any fraction of the endowment to a public account, the remainder going to the subject's private account; (b) the total group contribution, the subject's gross earnings, and the contributions of other group members (presented in random order) were then revealed to each subject, who was then permitted to assign sanctions to others. Finally, payoffs were calculated according to (1), and subjects were informed of their net payoffs for the period. They were then again randomly reassigned to groups and the process continued.

Our experimental results confirmed the following:

Hypothesis 1: Punishing occurs whenever shirking occurs. Punishment occurs in all periods and under all treatment conditions when $a_i < w$ for some i. Indeed, 89% of our subjects exercised the punishment option at least once, and in no treatment was the fraction punishing less than 80%.

Hypothesis 2: The level of punishment directed toward player i increases with the cost imposes on individual punishers, $q(w - a_i)$.

Hypothesis 3: Shirkers respond to punishment. Punishment in one round leads shirkers to increase their contributions in subsequent rounds.

Hypothesis 4: Punishment Fosters Contributions. The level of contributions

does not decay when costly punishment is permitted.

Hypothesis 5: Altruism Does Not Explain Punishment. We will explain this result below.

Because we are interested in how behavior changes over time as players learn more about the consequences of their actions, we used the panel nature of our data to estimate a number of the implied learning models. A summary of our analysis is presented in Tables 1 and 2 of the paper cited above. It is possible that those punishing low contributors sought to modify the behavior of the shirkers in order to raise the payoffs of others. But were this the case subjects would both contribute more in larger groups (because for a given q, more benefits to others are distributed in large groups) and punish more in large groups (because if successful in inducing the free rider to contribute more it would generate more aggregate benefits.) The fact that group size *per se* has no effect on either punishment or contributions suggests that altruism toward other group members is not what is generating the high levels of punishment of free-riders.

A further test is the following. If our subjects correctly estimated the responsiveness of those punished in subsequent periods we can then calculate the degree of altruism which would have made punishment a best response given these beliefs. Could plausible levels of altruism explain the punishing behavior? The answer is no: in the smaller of our groups punishment actually lowers average benefits (the cost of the punishment is not made up by the subsequent higher contributions of those punished) so even if the punisher cared as much about others payoffs as his own, punishment would not "pay." We conclude that motives other than a concern of the payoffs of others motivates punishment.

While we think it likely that anger at low contributors was an important motive for punishment the role of emotions is more clearly revealed in the responses of the targets of punishment. Subjects responded to punishment in the following way. Those giving less than the mean ("shirkers") when punished contributed more, and the effect of punishment on contribution was larger the farther away from the mean. Those contributing more than the mean ("good citizens") also responded to punishment but in the opposite direction: good citizens did not revert to the mean unless they were punished, in which case they strongly reduced their contributions. These results are all statistically significant at conventional levels.

Is the shirkers' positive response to punishment a best response defined over the payoffs of the game? Or, by contrast, does shirking still pay even when the expected costs of punishment are considered? Our estimates indicated that shirkers receive sanctions of 0.25 points for each point not contributed to the group project and this punishment response to shirking appears not to vary across groups. The act of shirking deprives the shirker of the returns from the public project, so the net benefit of shirking in the absence of punishment is just $1 - q$. Comparing the

benefits of shirking $(1-q)$ with the cost (0.25) we find that for the two low-q groups shirking pays quite well (0.75-0.5) while for the high q groups it does not (0.3-0.5). Of course we do not know that the subjects correctly estimated the effect of shirking on the likelihood of being punished, but the econometric estimate is quite precise ($t = 12$) and it seems plausible that at least in the later rounds of the experiment subjects had an approximate idea of the punishment costs of shirking.

The conclusion is that responding positively to punishment is not a best response defined over the payoffs of the game. Our interpretation, which we develop in the next section, is that punishment signaled social disapproval which evoked an emotion of shame in the shirkers and they responded positively so as to relieve that uncomfortable feeling. A reasonable interpretation of good citizens' behavior is that group members respond spitefully to being punished only when it is clear they are contributing well above the norm.

This interpretation is consistent with the results of a remarkable public goods with punishment experiment implemented in 18 rural communities in Zimbabwe by Barr (2001). The game was structured along the above lines, except for the punishment stage, in which there was no option to reduce the payoffs to others. Rather, following the contribution stage, Barr's assistant would stand beside each player in turn and say "Player number __, Mr/Mrs __, contributed __. Does anyone have anything to say about that?" followed by an opportunity for all group members to criticize or praise the contributor. A quarter of the participants were criticized for contributing too little ("stingy," "mean," "Now I know why I never get offered food when I drop by your house!") Five percent were criticized for giving too much ("stupid," "careless with money"). Those who made low contribution and were criticized made larger contributions in subsequent rounds. Moreover, those who contributed a low amount and escaped criticism, but had witnessed the criticism of others who had contributed a similar amount, increased their contributions by *even more than those directly criticized*. As in our experiments, those who had contributed a large amount and were criticized reduced their contribution in subsequent rounds. Where low contributions escaped criticism entirely contributions fell in subsequent rounds.

A second experiment with both monetary and non monetary punishment (Masclet, Noussair, Tucker and Villeval 2001) yielded similar results with the interesting twist that the response to being awarded "punishment points" was significantly greater when they carried no monetary penalty than when they resulted in payoff reductions of the players. This was true in both a stranger and a partner treatment, but more so in the latter.

3 Reciprocity, Shame, and Punishment with Two Agents

Consider two agents who play a one-shot public goods game, and who (a) are *self-interested* and thus care about their personal material payoffs; (b) are generically *altrustic or spiteful* and thus place some weight, positive, negative, or zero, on the payoffs of the other players, independent from of their beliefs about the others' types or their past behavior; and (c) are *strong reciprocators* and thus, depending on the other's type, value their payoffs (positively or negatively); (d) have contribution norms, indicating to what extent it is ethically valuable to contribute, and if they violate their own norms, they experience *guilt*; and finally (e) experience *shame* if they violate their own personal values and are publicly sanctioned for this behavior. The altruism and strong reciprocity of these individuals may lead them to value the payoffs of others in the public goods game and thus to contribute on others' behalf. The strong reciprocity motive may lead the individual to engage in costly punishment of those contributing little (reducing their payoffs). Finally, anticipation of punishment, guilt may lead individuals to contribute.

We assume each agent starts with a personal account equal to 1 unit. Each agent contributes $a_i \in [0, 1]$, and then each receives $q(a_1+a_2)$, where $q \in (1/2, 1)$. Thus, the agents do best when each cooperates ($a_i = 1$), but each has an incentive to defect ($a_i = 0$) no matter what the other does. At the end of this *production period* there is a second period, which we call the *punishment period*, in which the agents are informed of the contribution of the other agents, and each agent $i = 1, 2$ may impose a penalty μ_{ij} on the other agent ($j \neq i$) at a cost $c(\mu_{ij})$. For illustrative purposes, we will assume $c(\mu) = \gamma\mu^2/2$.

The material payoffs to the agents are thus given by

$$\pi_i = 1 - a_i + q(a_1 + a_2) - \mu_{ji}, \tag{2}$$

plus the cost of punishing j, which is $\gamma\mu_{ij}^2/2$, where $j \neq i$. We have not included the last expression in i's material payoff for reasons explained below (in fact, simulations show that this choice does not affect the general behavior of the model). In each equation, the first two terms give the amount remaining in the agent's private account after contributing, the third term is the agent's share of the total reward from cooperation and the fourth term is the punishment inflicted upon the agent.

We assume each player i suffers a psychic cost $\beta_i(a_i^* - a_i)^2$ when he contributes a_i and his *contribution norm* is a_i^*. The parameter β_i thus measures the strength of the player's *guilt* at not living up to his ideals. It may seem odd that the agent is guilty if he contributes more than his ideal. But if that which he retains $(1 - a_i)$ is directed to other "worthy" purposes about which he also has norms, then the symmetry of guilt around a_i^* becomes reasonable.

We represent the weight that agent i places on the material payoff π_j Of agent $j \neq i$ by

$$\delta_{ij} = \alpha_i + \lambda_i(a_j - a_i^*) \tag{3}$$

The parameter α_i reflects the agent's *unconditional altruism* motive towards the other players. We assume $\alpha_i \geq 0$ (benevolence) in this illustrative model, but in general $\alpha_i < 0$ (spite) is also possible. The parameter $\lambda_i \geq 0$ is the agents *reciprocity* motive. Note that when $\lambda_i > 0$, i is more favorably inclined towards j, the larger is j's contribution compared to i's contribution norm a_i^*. Finally, note that we do not apply δ_{ij} to j's cost of punishing i, because we consider it implausible that i will increase his contribution because he cares about j and he realizes that j will have to punish him if he contributes too little.

We include the shame s_i experienced by agent i by including negatively in the utility function the psychic costs of being punished:

$$s_i = \sigma_i(a_i^* - a_i)\mu_{ji}, \tag{4}$$

where $j \neq i$, and the contribution norm a_i^* is a function of the endowments $w_1, \ldots, w_n$. Thus σ_i is a measure of the susceptibility of agent i to feeling shame. Note that the shame term is positive only if one has contributed less than one's contribution norm. Otherwise this term represents spite, since in this case when an agent is punished, lowering his contribution increases his utility.

Thus, the objective functions of the two agents are given by

$$u_i = \pi_i + \delta_{ij}\pi_j - \beta_i(a_i^* - a_i)^2 - \gamma\mu_{ij}^2 - \sigma_i(a_i^* - a_i)\mu_{ji} \tag{5}$$

where $j \neq i$. Note that each agent i must choose a_i, and then choose μ_{ij} as a function of the level of contribution a_j chosen by the other agent. The first order condition for $\mu_{ij}(j \neq i)$ is given by

$$\frac{\partial u_i}{\partial \mu_{ij}} = -\gamma\mu_{ij} - \delta_{ij} = 0, \tag{6}$$

where $j \neq i$. This requires that the agent choose a level of punishment that equates the marginal cost of punishment (the first term) to the marginal benefit of punishment, namely the valuation placed on reducing the payoff of the other (the second term). This has the solution

$$\mu_{ij} = \begin{cases} -\delta_{ij}/\gamma & a_i^* > a_j + \alpha_i/\lambda_i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $j \neq i$. Where punishment is positive, this is clearly increasing in the degree or reciprocity and decreasing in the level of altruism.[4]

[4]Special cases not included in this solution are: if $\lambda_i = 0, \alpha_i \geq 0$, then $\mu_{ij} = 0$, and if $\lambda_i = 0, \alpha_i < 0$, then $\mu_{ij} = -\alpha_i/\gamma$. We will assume that if $\lambda_i = 0$ then the agent is purely selfish, so $\alpha_i = 0$ also holds.

We assume that each player i, in selecting a contribution level a_i, knows (7), and thus anticipates the effect of contributing more on the punishment one may expect to receive from another player j. The first order condition for a_i, $\partial u_i/\partial a_i = 0$, is then given by

$$\frac{\partial u_i}{\partial a_i} = -1 + q + \frac{\lambda_j}{\gamma} + \delta_{ij} q + 2\beta_i(a_i^* - a_i) + \sigma_i \left(\mu_{ji} + (a_i^* - a_i)\frac{\lambda_j}{\gamma} \right) = 0. \quad (8)$$

Note that $1-q$ is the marginal cost of contributing, λ_j/γ is the marginal reduction in punishment associated with contributing more, $\delta_{ij} q$ is the valuation of the marginal effect of contributing on the payoffs of the other agent, $2\beta_i(a_i^* - a_i)$ is the marginal reduction in guilt and the last two terms are the reduction in shame occasioned by the lesser violation of one's norm (the penultimate term), and the reduced punishment (the final term).

Equation (8) gives the best response function

$$a_i = \frac{\gamma(q(1+\delta_{ij}) + 2\beta_i a_i^* - 1) + \sigma_i(\lambda_j(a_j^* + a_i^*) - \alpha_j) + \lambda_j}{2(\beta_i\gamma + \lambda_j\sigma_i)} \quad (9)$$

For positive levels of reciprocity, the best response of each individual is increasing in the contribution of the other. Comparative static analysis of each shows that for $a_i^* > a_i$, $da_i/d\sigma_i$ and $da_i/d\beta_i$ are both positive, so an increase in either guilt or shame shifts the relevant function upwards. Figure 2 presents the best response functions for the two individuals, their intersection giving the Nash equilibrium. The shifts in the best response functions in figure 2 illustrate the effect on the Nash equilibrium of an increase in shame by individual 1 and an increase in guilt by individual 2. The effects of both, singly and together, are to increase the equilibrium contributions of both individuals.

Solving the resulting set of best response functions to get a Nash equilibrium is straightforward, and the equilibrium is unique. The expression for the solution is complicated, however, and we will not list it here.

We can also give some comparative static results for particular ranges of the parameters. First, suppose the two agents have the same behavioral attributes, so $\alpha_1 = \alpha_2$, $\lambda_1 = \lambda_2$, $\beta_1 = \beta_2$, $\sigma_1 = \sigma_2$, and $\alpha_1^* = a_2^*$. Furthermore, suppose $\gamma = 2$, $q = 0.75$, $a_1^* = 0.7$, $\alpha_1 = 0.01$, $\lambda_1 = 0.2$, $\beta_1 = 0.5$, and $\sigma_1 = 0.5$. Then we find $a_1 = a_2 = 0.55$, and each agents punishes the other a small amount, $\mu_{ij} = 0.01$. For the comparative statics, let us first vary σ_1 from zero to 30. We find that the equilibrium contribution increases from 0.54 to about 0.66. The reason for this small effect of shame is that the guilt parameter $\beta_1 = 0.5$ is rather large. If we reduce this to $\beta_1 = 0.20$, then the equilibrium contribution increases from 0.15to 0.65 when we vary σ_1 from zero to 30. The central point is that simulations show

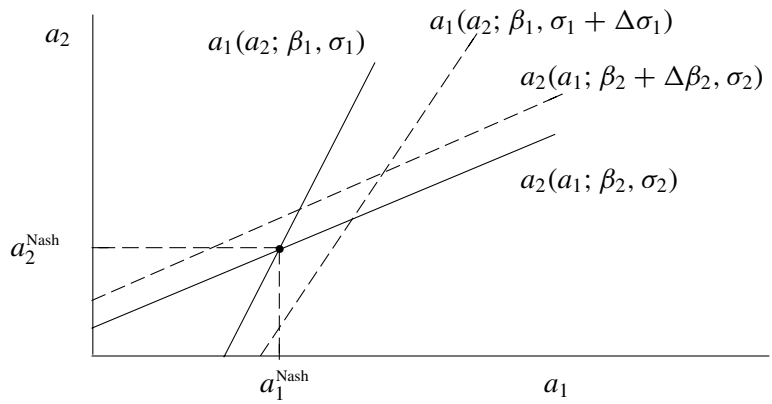

Figure 2: Comparative Statics. The best response functions (9) determine the Nash equilibrium contribution levels (a_1^{Nash} and a_2^{Nash}), which are both displaced upwards by an increase in 1's level of shame and or 2's level of guilt.

that an increase in shame leads to an increase in cooperation, and to a decline in the amount of punishment meted out.

The guilt parameter behaves similarly. If we increase β_i from 0.17 to 2, equilibrium contribution increases from zero to 0.64, and equilibrium punishment declines from 0.045 to zero.

We can similarly show that increasing the altruism parameter α_1, the reciprocity parameter λ_1, or the contribution norm a_1^* leads to an increase in the amount of cooperation.

Second, suppose one player is as before, but the second is perfectly self-interested, with $\alpha_2 = \lambda_2 = \beta_2 = \sigma_2 = a_2^* = 0$. In this case, player 2 never punishes, first order condition for player 1 is given by

$$\frac{\partial u_1}{\partial a_1} = -1 + q(1+\delta_{12}) + 2\beta_1(a_1^* - a_1) = 0, \tag{10}$$

while for player 2 we have

$$\frac{\partial u_2}{\partial a_2} = -1 + q + \frac{\lambda_1}{\gamma}. \tag{11}$$

Thus if $\lambda_1/\gamma > 1 - q$, player 2 will set a_2 to the value that equates μ_{12} to zero (since, by definition, μ_{12} cannot be negative). This gives

$$a_2 = a_1^* - \frac{\alpha_1}{\lambda_1}, \tag{12}$$

and

$$a_1 = a_1^* - \frac{1-q}{2\beta_1}. \tag{13}$$

In this case, which occurs when the cost of punishing, γ, is low and the intensity of reciprocation, λ_1 is high, is relatively efficient, since no punishment actually occurs. Nevertheless, the agents do not attain their contribution norms, and the outcome could be far from optimal if the reciprocator's contribution norm is low.

Conversely, if $\lambda_1/\gamma < 1-q$, player 2 will set $a_2 = 0$, so

$$a_1 = a_1^*\left(1 - \frac{q\lambda_1}{2\beta_1}\right) - \frac{1-q(1+\alpha_1)}{2\beta_1}. \tag{14}$$

The level of punishment of the self-interested player is given by

$$\mu_{12} = \frac{\lambda_1 a_1^* - \alpha_1}{\gamma},$$

which is bounded above by $(1-q)a_1^*$. Thus there can be extensive punishment in this case, although it does not induce the selfish type to cooperate. Also, it is clear that the level of punishment is increasing in the reciprocator's contribution norm, a_1^*, the intensity of reciprocation, λ_1, and is decreasing in the reciprocator's level of altruism, α_1, and the cost of punishment, γ.

We can also show, using this asymmetric model, that if agent 2 moves from being purely self-interested to experiencing shame, the average level of contribution of the agents will increase and the amount of punishment will decline. So let us now suppose that $\sigma_2 > 0$, $\alpha_2^* = a_1^*$, and $1-q > \lambda_1/\gamma$, so agent 2 contributes nothing when $\sigma_2 = 0$. We also assume $a_1^* > \alpha_1/\lambda_1$, without which no punishment can occur. In this case the equilibrium value of a_2 is

$$a_2 = \frac{\sigma_2(2\lambda_1 a_1^* - \alpha_1) + \lambda_1 - (1-q)\gamma}{2\lambda_1\sigma_2}.$$

This expression is negative when $\sigma_2 = 0$, as we would expect. But a_2 is increasing in σ_2, and is positive for sufficiently large σ_2. The amount punishment is

$$\mu_{12} = \frac{\lambda_1(a_1^* - a_2) - \alpha_1}{\gamma},$$

so clearly punishment declines as agent 2's shame level increases. Finally, we have

$$a_1 = a_1^* - \frac{(1-q) - q(\alpha_1 - \lambda_1(a_2 - a_1^*))}{2\beta_1},$$

which is clearly increasing in a_2. Hence when agent 2's shame level increases, both players contribute more and the level of punishment declines.

We now develop a more general model of cooperation in a public goods game in which individuals have the same structure of preferences as in the previous section.

4 A General Model of Reciprocity, Shame, and Punishment

Consider a group with members $i = 1, \ldots, n$, each of whom has an endowment w_i and can make a contribution $a_i \in [0, w_i]$ that yields a payoff $f(a_1, \ldots, a_n)$ to each member of the group, where f is increasing in each of its arguments, but $\partial f/\partial a_i < 1$, so a member does best contributing as little as possible, everything else being equal. At the end of this *production period* there is a second period, which we call the *punishment period*, in which each member i of the group is informed of the vector of endowments $(w_1, \ldots, w_n)$ and contributions $(a_1, \ldots, a_n)$ and imposes a penalty μ_{ij} on each member $j \neq i$ at a cost $c_i(\mu_{ij})$ to himself. For notational convenience, we assume $\mu_{ii} = 0$ and $c_i(0) = 0$. We define the *material payoff* to member i as

$$\pi_i = w_i - a_i + f(a_1, \ldots, a_n) - \sum_{j \neq i} \mu_{ji}. \tag{15}$$

For the reason described in the previous section, we have not included the cost to i of punishing others in the expression for π_i.

The fact that agents punish other agents in this model flows from the assumption that agents are not entirely self-regarding. Rather, they place a positive weight on the payoffs of other agents whom they favor, and a negative weight on other agents whom they disfavor. Agent i's assessment of j's type is a function of a_j. Generalizing the two-person model, the weight δ_{ij} that i places on j's material payoff is given by (3), the disutility of shame is given by (4), and the psychic cost of guilt is $\beta_i(a_i^* - a_i)^2$. The utility of member i is then given by

$$u_i = \pi_i + \sum_{j \neq i} \delta_{ij}\pi_j - \sum_{j \neq i} c_i(\mu_{ij}) - \beta_i(a_i^* - a_i)^2 - \sigma_i(a_i^* - a_i) \sum_{j \neq i} \mu_{ji}, \tag{16}$$

The first order condition for μ_{ij}, $j \neq i$, is given by

$$\frac{\partial u_i}{\partial \mu_{ij}} = -\frac{\partial c_i}{\partial \mu_{ij}} - \delta_{ij} \leq 0, \tag{17}$$

and equality holds if $\mu_{ij} > 0$. Assuming equality in the first order condition, and totally differentiating with respect to a_j, we have

$$\frac{\partial^2 u_i}{\partial \mu_{ij}^2} \frac{d\mu_{ij}}{da_j} + \frac{\partial^2 u_i}{\partial \mu_{ij} \partial a_j} = 0.$$

The first double partial is negative by the second order condition, and $\frac{\partial^2 u_i}{\partial \mu_{ij} \partial a_j} = -\lambda_i < 0$. Hence $\frac{d\mu_{ij}}{da_j} < 0$, which means that when j's contribution increases, i punishes j less (or at least not more).

The first order condition for a_i is given by

$$\frac{\partial u_i}{\partial a_i} = -1 + \frac{\partial f}{\partial a_i}(1 + \sum_{j \neq i} \delta_{ij}) + 2\beta_i(a_i^* - a_i)$$
$$+\sigma_i \sum_{j \neq i} \mu_{ji} - \sum_{j \neq i} (\delta_{ij} + \sigma_i(a_i^* - a_i)) \frac{\partial \mu_{ji}}{\partial a_i} = 0. \tag{18}$$

Totally differentiating the first order conditions with respect to σ_i, we get

$$\mathbf{J} \begin{bmatrix} \frac{da_1}{d\sigma_i} \\ \vdots \\ \frac{da_1}{d\sigma_i} \end{bmatrix} = - \begin{bmatrix} \frac{\partial^2 u}{\partial a_1 \partial \sigma_i} \\ \vdots \\ \frac{\partial^2 u}{\partial a_1 \partial \sigma_i} \end{bmatrix}.$$

where $\mathbf{J}$ is the Hessian matrix associated with the optimization. However, we have

$$\frac{\partial^2 u_i}{\partial a_i \partial \sigma_i} = -(a_i^* - a_i) \sum_{j \neq i} \frac{\partial \mu_{ji}}{\partial a_i},$$

and

$$\frac{\partial^2 u_i}{\partial a_j \partial \sigma_i} = 0, \qquad \text{for } j \neq i.$$

Thus

$$\begin{bmatrix} \frac{da_1}{d\sigma_i} \\ \vdots \\ \frac{da_1}{d\sigma_i} \end{bmatrix} = -\mathbf{J}^{-1} \begin{bmatrix} 0 \\ \vdots \\ (a_i^* - a_i) \sum_{j \neq i} \frac{\partial \mu_{ji}}{\partial \sigma_i} \\ \vdots \\ 0 \end{bmatrix}.$$

If we solve for $da_i/d\sigma_i$ using Cramer's rule, and using the fact that the second order condition for a maximum requires that the determinant of $\mathbf{J}$ and that of the ith principal minor must have opposite sign, we conclude that $da_i/da_i^* > 0$ when $a_i^* > a_i$; i.e., *if an agent is contributing less than his contribution norm, an increase in the strength of shame will induce the agent to contribute more.* It is precisely in this sense that shame is a prosocial emotion. The same reasoning shows that an individual who is contributing more that he thinks morally warranted (perhaps to avoid being punished), and he is punished anyway, he will respond by *reducing* his contribution when the shame-spite factor σ_i is increased.

Totally differentiating the first order conditions with respect to α_i^*, we get

$$\mathbf{J} \begin{bmatrix} \frac{da_1}{da_i^*} \\ \vdots \\ \frac{da_1}{da_i^*} \end{bmatrix} = - \begin{bmatrix} \frac{\partial^2 u}{\partial a_1 \partial a_i^*} \\ \vdots \\ \frac{\partial^2 u}{\partial a_1 \partial a_i^*} \end{bmatrix}.$$

where **J** is the Hessian matrix associated with the optimization. However, we have

$$\frac{\partial^2 u_i}{\partial a_i \partial a_i^*} = 2\beta_i - \sigma_i \sum_{j \neq i} \frac{\partial \mu_{ji}}{\partial a_i},$$

and

$$\frac{\partial^2 u_i}{\partial a_j \partial a_i^*} = 0, \qquad \text{for } j \neq i.$$

Thus

$$\begin{bmatrix} \frac{da_1}{da_i^*} \\ \vdots \\ \frac{da_1}{da_i^*} \end{bmatrix} = -\mathbf{J}^{-1} \begin{bmatrix} 0 \\ \vdots \\ 2\beta_i - \sigma_i \sum_{j \neq i} \frac{\partial \mu_{ji}}{\partial a_i} \\ \vdots \\ 0 \end{bmatrix}.$$

If we solve for da_i/da_i^* using Cramer's rule, and using the fact that the second order condition for a maximum requires that the determinant of **J** and that of the ith principal minor must have opposite sign, we conclude that $da_i/da_i^* > 0$; i.e., *if an agent raises his contribution norm, his contribution increases.*

5 The Bioeconomics of Prosocial Emotions

The Adam Smith of *The Theory of Moral Sentiments* is of course much less well known and less studied than the Adam Smith of *The Wealth of Nations.* Generations of economists have puzzled that the same Scottish philosopher whose analysis of emotion is perhaps the greatest in the English language before William James (1884), could also give us an even more famous discourse based on the idea that "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest." In act, we now know from laboratory experiments that subjects in market-like situations behave like the Adam Smith of *The Wealth of Nations*, while their behavior in strategic interactions resembles more the Adam Smith of *The Theory of Moral Sentiments.* No doubt this is the distinction Smith had in mind in writing his two great books.

Economic theorists have long been happy to assume that individuals act to maximize their private gain. While arguments in favor of this assumption are rarely systematically presented, the informal argument is that other types of behavior should be driven from the scene by the relentless success of the self-interested types. This argument may be plausible when profit-oriented firms are the object of analysis, but why should it hold when subjective utility is the object of our strivings?

To answer this question, arguments from biology have conveniently stepped in to fill the breech (Hamilton 1964, Williams 1966, Trivers 1971, Maynard Smith and

Price 1973). Evolution ensures that only the self-interested survive. What appears to be altruism— personal sacrifice on behalf of others—is really just self-interest at the genetic level. Richard Dawkins (1989), for instance, writes "We are survival machines—robot vehicles blindly programmed to preserve the selfish molecules known as genes.…This gene selfishness will usually give rise to selfishness in individual behavior." Similarly, in a famous work devoted exclusively to human sociality, R. D. Alexander (1987) asserts that "ethics, morality, human conduct, and the human psyche are to be understood only if societies are seen as collections of individuals seeking their own self-interest." (p. 3).

While the empirical evidence shows that humans systematically deviate from the model of the self-interested actor, and we think the evidence is strong that prosocial emotions account for much of this behavior, we must remain unsatisfied with alternative descriptions of behavior until we understand how prosocial emotions might have evolved, culturally, genetically, or both, and what forces prevent the deterioration of nonselfish preferences once they are established. The puzzle here is that prosocial emotions are at least *prima facie* altruistic, benefiting others at a cost to oneself, so that under simple replicator dynamics, in which the selfishly favorable trait tends to increase in frequency, prosociality should atrophy. This question is, of course, the subject of active research these days among economists and other decision theorists (Frank, 1987, 1988; Eckman, 1992; Damasio, 1994; Elster, 1998). We will not propose a definitive answer, but rather suggest some fruitful lines of research and the reasoning on which they are based.

5.1 The Internalization of Norms

One does not feel shame merely because one is thought ill of by one's social group. Indeed, if one has acted honorably according to one's own values, and one is nevertheless punished, one feels spiteful rather than shameful. This is indicated in our model by the fact that the sign of the shame term depends on whether $a^* > a$, in which case one feels shame when punished, and hence acts to increase one's contribution a, or $a^* > a$, in which case one feels spite, and hence acts to decrease one's contribution a. The parameter a^* is thus a personal attribute that is absolutely central to how one reacts emotionally to group sanctions. What sort of entity is α^*?

Parameter α^* is an *internalized norm*. In general, a *norm* is a rule sanctioned by a group and followed by its members (Axelrod 1986, Elster 1989, Jordan 1991, Frank 1991, Kandori 1992, Boyd and Richerson 1994, Binmore and Samuelson 1994, Bowles and Gintis 1998b, Henrich and Boyd 2001). Generally, then, norms are *constraints* that one must obey in maximizing one's welfare (e.g., the norm of honesty in commercial transactions), presumably because violating the norm

would be more costly than obeying it. An *internalized* norm is a norm that one has accepted, not as a constraint, but rather as an *argument of one's objective function*. We strive to conform to internalized norms not because we will be punished if we do not conform, but because we actively *wish* to conform. For instance, consider the norm of 'helping individuals in distress.' I may help an individual in distress because I will be rewarded by my social group for doing so, or I will be punished for not doing so. If the norm is internalized, however, I help because I personally and genuinely want to (or at least believe I should want to), and if I did not help, I would feel *guilt*. Moreover, if I were discovered not helping, I would feel *shame*. In the latter case, I have 'internalized' the norm of helping people in distress.

Sociological theory treats the internalization of norms as a central element the analysis of prosocial behavior (Durkheim 1951, Boas 1938, Benedict 1934, Mead 1963, Geertz 1963, Parsons 1967, Grusec and Kuczynski 1997). Norms are internalized from parents (*vertical transmission*), influential elders and insitutional practices(*oblique transmission*), and one's peers (*horizontal transmission*) (Cavalli-Sforza and Feldman 1981, Boyd and Richerson 1985). The psychological mechanisms that account for internalization are doubtless complex, and the phenomenon is probably unique to our species. The fully informed, self-interested optimizer of standard economic theory would not internalize a norm, since doing so places value on the norm above and beyond the extrinsic social benefit of conforming to it and social cost of violating it, so the opimizer will conform more closely to the norm than he would if he treated it simply as a constraint. So why does internalization exist?

The answer is that human society is so complex and the benefits and costs of conforming to or violating its many norms so difficult to assess, that full-scale optimization using norms as constraints is excessively, and even perhaps fatally, error-prone. The internalization of norms eliminates many of the cost/benefit calculations and replaces them with simple moral and prudential guidelines for action. Individuals who internalize norms are therefore more biologically fit than those who do not, so the psychological mechanisms of internalization are evolutionarily selected.

There are two important implications of norm internalization for the economic analysis of social cooperation. The first is that when an agent internalizes a norm, it remains an argument in his utility function in *all* social settings. This explains why an individual who has a norm of 'rejecting low offers,' which serves him well in daily life by helping build a reputation for hard bargaining, will continue to embrace this norm in a one-shot ultimatum game. Norm internalization thus helps explain the otherwise anomalous behavior exhibited in laboratory bargaining settings. Doubtless other internalized norms help explain behavior in public goods, trust, dictator, and the other strategic setting used in behavioral game theory.

The second important implication of norm internalization is that it can explain *altruistic* behavior, in which the individual behaves in a way that is personally costly but that benefits the group—as, for instance, punishing noncontributors in a public goods game. The connection between altruism and internalization was first proposed by Herbert Simon (1990), who suggested that if internalization (Simon called it 'docility,' in its root meaning of 'easy to mold or shape') is in general fitness enhancing, then social institutions could add to the set of norms transmitted vertically and obliquely, some that in fact are fitness reducing for the individual, though group beneficial. Gintis (2003) provides an analytically rigorous genetic model demonstrating the plausibility of Simon's theory.[5]

Empirically, all societies indeed promote a combination of self-regarding and altruistic norms. All known cultures foster norms that enhance personal fitness, such as prudence, personal hygiene, and control of emotions, but also promote norms that subordinate the individual to group welfare, fostering such behaviors as unconditional bravery, honesty, fairness, and willingness to cooperate, to refrain from overexploiting a common pool resource, to vote and otherwise participate in the political life of the community, to act on behalf of one's ethnic or religious group, and to identify with the goals of an organization of which one is a member, such as a business firm or a residential community (Brown 1991). The central tenets of virtually all of the world's great religions also exhibit this tendency of combining personally fitness-enhancing norms and altruistic norms, as well as denying that there is any essential difference between the two.

One important social norms is 'reward those who obey social norms and punish those who do not.' This norm is clearly altruistic, and is subject to internalization. Those who internalize this norm in the public goods game are precisely those with high λ's.

5.2 Pain

Shame is one of the seven so-called "social" emotions, of which the others are love, guilt, embarrassment, pride, envy, and jealousy (Plutchik 1980, Eckman 1992). Shame has a similar role in regulating social behavior as does pain in regulating behavior in general, so we shall begin with an analysis of the role of pain in decision-making. Pain is one of the six so-called 'basic' emotions, the others being pleasure, anger, fear, surprise, and disgust. Basic and social emotions are universally expressed in human societies, although their expression is affected by cultural

[5]The central problem any such model must handle is why those who internalize both the fitness-enhancing norms and the altruistic norms are not out-competed by those who internalize only the fitness-enhancing norms. An analysis of genotype-phenotype interaction explains why this 'unraveling' of altruistic behavior need not occur.

conditions. For instance, one may be angered by an immoral act, or disgusted by an unusual foodstuff, but what counts as an immoral act or a disgusting foodstuff is, at least to some extent, culturally specific.

Even the simplest forms of life have some way to affect their local environment. One-celled organisms, such as Euglena and Paramecium, for instance, have flagella and cilia of marvelous construction that allow the creature to locomote. Most simple creatures move in reaction to various temperature, pressure, or chemical gradients, or simply move randomly when local conditions are poor for survival. More complex organisms have the ability to repair damage to themselves, and to learn to avoid such damage in the future. In humans and many other vertebrates this takes the form of *pain*—a highly aversive sensation that the organism simply cannot ignore and will do virtually anything to avoid in the future.

Yet an organism with complete information, an unlimited capacity to process information, and with an fitness-maximizing way of discounting future costs and benefits would have no use for pain. Such an agent would be able to assess the costs of any damage to itself, would calculate an optimal response to such damage, and would prepare optimally for future occurrences of this damage. The aversive stimulus—pain—would then not simply be otiose. It would rather be strongly distorting of optimal behavior. Because pain will lead the agent to assuasive and avoidance behavior *in addition to* responding constructively to the damage. Since pain clearly does have adaptive value, it follows that modeling pain *presupposes* that the agent experiencing pain must have incomplete information and/or a limited capacity to process information, and/or an excessively high rate of discounting future benefits and costs.

5.3 Shame

Pain is a pre-social emotion. Shame is a social emotion: a distress that is experienced when one is devalued in eyes of one's consociates because of a value that one has violated or a behavioral norm that one has not lived up to.

Does shame serve a purpose similar to that of pain? If being socially devalued has fitness costs, and if the amount of shame is closely correlated with the level of these fitness costs, then the answer is affirmative. Shame, like pain, is an aversive stimulus that leads the agent experiencing it to repair the situation that led to the stimulus, and to avoid such situations in the future. Shame, like pain, replaces an involved optimization process with a simple message: whatever you did, undo it if possible, and do not do it again.

Since shame is evolutionarily selected and is costly to use, it must confer a selective advantage on those who experience it. Two types of selective advantage

are at work here. First, shame may raise the fitness of an agent who has incomplete information (e.g., as to how fitness-reducing a particular anti-social action is), limited or imperfect information-processing capacity, and/or a tendency to undervalue costs and benefit that accrue in the future. Probably all three conditions conspire to react suboptimally to social disapprobation in the absence of shame, and shame brings us closer to the optimum. Of course the role of shame in alerting us to negative consequences in the future presupposes that society is organized to impose those costs on rule violators. The emotion of shame may have coevolved with the emotions motivating punishment of antisocial actions (the reciprocity motive in our model).

The second selective advantage to those experiencing shame arises through the effects of group competition. Where the emotion of shame is common, punishment of antisocial actions will be particularly effective and as a result seldom used .Thus groups in which shame is common can sustain high levels of group cooperation at limited cost and will be more likely to spread through interdemic group selection (Bowles and Gintis 1998a, Boyd, Gintis, Bowles and Richerson 2001). Shame thus serves as a means of economizing on costly within group punishment.

6 Conclusion

The experimental evidence and reasoning presented here suggest that there is something fundamentally wrong with the behavioral assumptions underlying the canonical approach to economic policy and constitution-making. This approach assumes that agents will maximize a pre-given objective function subject to whatever costs and benefits are defined by the policy or law. However, when agents consider the policy-making body in some appropriate sense valid and legitimate, they will avoid violating the rules on principle, and not only because they will be rewarded for doing so, or punished for transgressing the rules. Albert Hirschman (1985):10 described the situation this way:

> *Economists often propose to deal with unethical or antisocial behavior by raising the cost of that behavior rather than proclaiming standards and imposing prohibitions and sanctions. The reason is probably that they think of citizens as consumers with unchanging or arbitrarily changing tastes in matters of civic as well as commodity-oriented behavior. .A principal purpose of publicly proclaimed laws and regulations is to stigmatize antisocial behavior and thereby to influence citizens' values and behavior codes.*

Hirschman believes that penalties imposed on miscreants affect behavior in two ways: they alter the payoff consequences of various actions and they affect the

preferences that actors use in evaluating the consequences of their actions. His point is that economists are remiss in focusing entirely on the first. This narrow focus is nowhere more clear than in modern implementation theory, which seeks to design policies such that agents' given preferences lead them individually to act in ways that implement a socially valued outcome as a Nash equilibrium.

Hirschman is arguing against a venerable tradition, not only in economics, but in political philosophy as well, one dating back before Smith wrote his *Theory of Moral Sentiments*. In 1754, David Hume (Hume 1898[1754]) advised "that, in contriving any system of government…every man ought to be supposed to be a knave and to have no other end, in all his actions, than his private interest." But he was appealing to prudence, not to realism. His next sentence reads: it is "strange that a maxim should be true in politics which is false in fact." However if, as Hume realized, individuals are not uniformly selfish, but rather are sometimes given to the honorable sentiments about which Smith wrote, then prudence might recommend an alternative dictum: policy makers and constitution builders should know that populations are heterogeneous and the individuals making them up are both versatile and plastic, and that good policies and constitutions are those that support socially valued outcomes not only by harnessing selfish motives to socially valued ends, but also by evoking, cultivating, and empowering public spirited motives. It is not as tidy as Hume's dictum, and implementing it requires the analysis of the emergent properties of rather complex interactions among heterogeneous agents, but both realism and prudence may be claimed for it.

References

Alexander, R. D., *The Biology of Moral Systems* (New York: Aldine, 1987).

Andreoni, James, "Cooperation in Public Goods Experiments: Kindness or Confusion," *American Economic Review* 85,4 (1995):891–904.

Arrow, Kenneth J., "Political and Economic Evaluation of Social Effects and Externalities," in M. D. Intriligator (ed.) *Frontiers of Quantitative Economics* (Amsterdam: North Holland, 1971) pp. 3–23.

Axelrod, Robert, "An Evolutionary Approach to Norms," *American Political Science Review* 80 (1986):1095–1111.

— and William D. Hamilton, "The Evolution of Cooperation," *Science* 211 (1981):1390–1396.

Barr, Abigail, "Social Dilemmas, Shame Based Sanctions, and Shamelessness: Experimental results from Rural Zimbabwe," 2001. Oxford University.

Benedict, Ruth, *Patterns of Culture* (Boston: Houghton Mifflin, 1934).

Binmore, Ken and Larry Samuelson, "An Economists Perspective on the Evolution of Norms," *Journal of Institutional and Theoretical Economics (*1994):45–63.

Boas, Franz, *General Anthropology* (Boston: Heath, 1938).

Bosman, Ronald and Frans van Winden, "Anticipated and Experienced Emotions in an Investment Experiment," 2001. Unpublished, Amsterdam.

Bowles, Samuel and Herbert Gintis, "The Evolution of Strong Reciprocity," 1998. Santa Fe Institute Working Paper #98-08-073E.

— and — , "The Moral Economy of Community: Structured Populations and the Evolution of Prosocial Norms," *Evolution & Human Behavior* 19,1 (January 1998):3–25.

— and — , "The Evolution of Human Sociality," 2001. Unpublished Book Manuscript.

— , Jeffrey Carptenter, and Herbert Gintis, "Mutual Monitoring in Teams: The Importance of Shame and Punishment," 2001. University of Massachusetts.

Boyd, Robert and Peter J. Richerson, *Culture and the Evolutionary Process* (Chicago: University of Chicago Press, 1985).

— and — , "The Evolution of Cooperation," *Journal of Theoretical Biology* 132 (1988):337–356.

— and — , "The Evolution of Norms: An Anthropological View," *Journal of Institutional and Theoretical Economics* 150,1 (1994):72–87.

— , Herbert Gintis, Samuel Bowles, and Peter J. Richerson, "Altruistic Punishment in Large Groups Evolves by Interdemic Group Selection," 2001. Working Paper.

Brown, Donald E., *Human Universals* (New York: McGraw-Hill, 1991).

Camerer, Colin and Richard Thaler, "Ultimatums, Dictators, and Manners," *Journal of Economic Perspectives* 9,2 (1995):209–219.

Cavalli-Sforza, Luigi L. and Marcus W. Feldman, *Cultural Transmission and Evolution* (Princeton, NJ: Princeton University Press, 1981).

Cosmides, Leda and John Tooby, "Cognitive Adaptations for Social Exchange," in Jerome H. Barkow, Leda Cosmides, and John Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (New York: Oxford University Press, 1992) pp. 163–228.

Croson, Rachel, "Partners and Strangers Revisited," *Economic Letters* 53 (1996):25–32.

Damasio, Antonio R., *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Avon Books, 1994).

Dawes, Robyn M., John M. Orbell, and J. C. Van de Kragt, "Organizing Groups for Collective Action," *American Political Science Review* 80 (December 1986):1171–1185.

Dawkins, Richard, *The Selfish Gene, 2nd Edition* (Oxford: Oxford University Press, 1989).

Durkheim, Emile, *Suicide, a Study in Sociology* (New York: Free Press, 1951). Translated by John A. Spaulding and George Simpson. Edited, with an Introduction by George Simpson.

Eckman, Paul, "An Argument for Basic Emotions," *Cognition and Emotion* 6 (1992):169–200.

Elster, Jon, "Social Norms and Economic Theory," *Journal of Economic Perspectives* 3,4 (1989):99–117.

— , "Emotions and Economic Theory," *Journal of Economic Perspectives* 36 (1998):47–74.

Falk, Armin and Urs Fischbacher, "A Theory of Reciprocity," 1998. Unpublished Manuscript, Institute for Empirical Economic Research, University of Zurich.

Fehr, Ernst and Klaus M. Schmidt, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics* 114 (August 1999):817–868.

— and Simon Gächter, "Cooperation and Punishment," *American Economic Review* 90,4 (September 2000):980–994.

Frank, Robert, "Social Forces in the Workplace," in Kenneth Koford and Jeffrey Miller (eds.) *Social Norms and Economic Institutions* (Ann Arbor, MI: University of Michigan Press, 1991) pp. 151–179.

Frank, Robert H., "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?," *American Economic Review* 77,4 (September 1987):593–604.

— , *Passions Within Reason: The Strategic Role of the Emotions* (New York: Norton, 1988).

Fudenberg, Drew and Eric Maskin, "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica* 54,3 (May 1986):533–554.

Geanakoplos, John, David Pearce, and Ennio Stacchetti, "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1 (March 1989):60–79.

Geertz, Clifford, *Peddlers and Princes: Social Change and Economic Modernization in Two Indonesian Towns* (Chicago: University of Chicago Press, 1963).

Gintis, Herbert, "The Hitchhiker's Guide to Altruism: Genes and Culture, and the Internalization of Norms," *Journal of Theoretical Biology (*2003).

Gneezy, Uri and Aldo Rustichini, "A Fine is a Price," *Journal of Legal Studies* 29 (2000):1–17.

Grusec, Joan E. and Leon Kuczynski, *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory* (New York: John Wily & Sons,

1997).

Güth, Werner and Reinhard Tietz, "Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results," *Journal of Economic Psychology* 11 (1990):417–449.

Hamilton, W. D., "The Genetical Evolution of Social Behavior," *Journal of Theoretical Biology* 37 (1964):1–16,17–52.

Henrich, Joseph and Robert Boyd, "Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas," *Journal of Theoretical Biology* 208 (2001):79–89.

Hirschman, Albert, "Against Parsimony," *Economic Philosophy* 1 (1985):7–21.

Hume, David, *Essays: Moral, Political and Literary* (London: Longmans, Green, 1898[1754]).

James, William, "What is an Emotion?," *Mind* 9 (1884):188–205.

Jordan, J. S., "Bayesian Learning in Normal Form Games," *Games and Economic Behavior* 3 (1991):60–81.

Kandori, Michihiro, "Social Norms and Community Enforcement," *Review of Economic Studies* 57 (1992):63–80.

Laibson, David, "A Cue-Theory of Consumption," 1996. Harvard University.

Levine, David K., "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics* 1,3 (1998):593–622.

Loewenstein, George F., "Out of Control: Visceral Influences on Behavior," *Organizational Behavior and Human Decision Processes* 65 (1996):272–292.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval, "Monetary and Non Monetary Punishemnt in the Voluntary Contributions Mechanism," 2001. Purdue University.

Maynard Smith, John and G. R. Price, "The Logic of Animal Conflict," *Nature* 246 (2 November 1973):15–18.

Mead, Margaret, *Sex and Temperament in Three Primitive Societies* (New York: Morrow, 1963).

Ostrom, Elinor, James Walker, and Roy Gardner, "Covenants with and without a Sword: Self-Governance Is Possible," *American Political Science Review* 86,2 (June 1992):404–417.

Parsons, Talcott, *Sociological Theory and Modern Society* (New York: Free Press, 1967).

Plutchik, R., *Emotion: A psychoevolutionary synthesis* (New York: Harper & Row, 1980).

Roth, Alvin, "Bargaining Experiments," in John Kagel and Alvin Roth (eds.) *The Handbook of Experimental Economics* (Princeton, NJ: Princeton University Press, 1995).

Sato, Kaori, "Distribution and the Cost of Maintaining Common Property Resources," *Journal of Experimental Social Psychology* 23 (January 1987):19–31.

Sethi, Rajiv and E. Somanathan, "Preference Evolution and Reciprocity," *Journal of Economic Theory* 97 (2001):273–297.

Simon, Herbert, *Models of Bounded Rationality* (Cambridge, MA: MIT Press, 1982).

— , "A Mechanism for Social Selection and Successful Altruism," *Science* 250 (1990):1665–1668.

Taylor, Michael, *Anarchy and Cooperation* (London: John Wiley and Sons, 1976).

Trivers, R. L., "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology* 46 (1971):35–57.

Williams, G. C., *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought* (Princeton, NJ: Princeton University Press, 1966).

Yamagishi, Toshio, "The Provision of a Sanctioning System in the United States and Japan," *Social Psychology Quarterly* 51,3 (1988):265–271.

— , "Seriousness of Social Dilemmas and the Provision of a Sanctioning System," *Social Psychology Quarterly* 51,1 (1988):32–42.

— , "Group Size and the Provision of a Sanctioning System in a Social Dilemma," in W.B.G. Liebrand, David M. Messick, and H.A.M. Wilke (eds.) *Social Dilemmas: Theoretical Issues and Research Findings* (Oxford: Pergamon Press, 1992) pp. 267–287.

Zajonc, R. B., "Feeling and Thinking: Preferences Need No Inferences," *American Psychologist* 35,2 (1980):151–175.