April 2006

Cruel to be kind: The role of the evolution of altruistic punishment in sustaining human cooperation in public goods games

Kelly L. Cataldo
University of Pennsylvania, kcataldo@sas.upenn.edu

Cruel to be kind: The role of the evolution of altruistic punishment in sustaining human cooperation in public goods games

Abstract
People cooperate in public goods games even when an individual's utility maximizing strategy is to defect. A form of non-institutionalized punishment called altruistic punishment—or strong reciprocity—may explain this cooperative behavior. I consider laboratory experiments of public goods games that provide evidence of altruistic punishment and proximate explanations for that behavior. I also present theories of the evolution of altruistic punishment via group-selection, multi-level selection, and gene and culture co-evolution. Furthermore, I consider criticisms of both laboratory results and evolutionary theories that suggest weaknesses in the current research on altruistic punishment. In sum, we will likely never have a definitive explanation of the origins and evolution of human cooperation. I conclude, however, that altruistic punishment may form an integral part of that trajectory.

Keywords
altruistic punishment, strong reciprocity, evolutionary game theory, human cooperation, Philosophy, Philos Politics & Econ, Brian Skyrms, Brian, Skyrms

Cruel to be kind: The role of the evolution of altruistic punishment in sustaining human cooperation in public goods games

Kelly L. Cataldo

PPE 475, Brian Skyrms

Submitted April 15, 2006

Table of Contents

Page

Introduction……..3

Laboratory Experiments: Evidence of Strongly Reciprocal Behavior……5

Proximate Explanations for Strongly Reciprocal Behavior……...11

Contested Explanations for the Evolution of Altruistic Punishment and Cooperation……14

The Evolution of Altruistic Punishment in Humans: Recent Theories of Group Selection, Multi-level Selection, and Gene and Culture Co-Evolution…….17

Criticisms of Group Selection Models……...…29

Discussion……..41

Conclusion……..……45

Endnotes…….46

References……...…47

Introduction

In modern human societies, people contribute to a variety of public goods, for example by paying taxes, even though the likelihood of being caught for non-participation is quite low. Yet the possibility for institutional punishment does exist. The contribution of so many individuals indicates that they believe that the threat of punishment is credible, albeit unlikely. Individuals also cooperate in daily interactions with unrelated others who they will probably never meet again. In the absence of an explicit punishment for non-cooperation, it is difficult to make sense of why people behave pro-socially. Some people will not. Certain individuals free-ride on the contributions of others, enjoying the benefits that cooperation provides without participating in the cooperative endeavor. Moreover, it makes sense for a utility-maximizing individual not to cooperate when cooperation is personally costly.

As in a public goods game, everyone is better off in any given society when people cooperate. When non-cooperation is the dominant strategy, there must be a mechanism that explains why we cooperate. The mechanism that I will explore is a form of non-institutionalized punishment called altruistic punishment. Altruistic punishment occurs when a cooperator punishes a defector. This punishment is altruistic because an individual punisher incurs a cost for sanctioning a non-cooperator. However, the reduction in payoffs that a defector sustains may deter her from defecting in similar situations in the future. Society is better off when individuals do not defect. Thus, this strategy may explain how cooperation can be sustained in the absence of institutionalized punishment.

I will begin with a discussion of laboratory experiments in which subjects altruistically punished others in public goods[i] games. I will focus exclusively on public goods games for two reasons. The public goods game is uniquely conducive to observing cooperative and punishing

behaviors in a social context. Because the public goods game "is designed to illuminate such behaviors as contributing to team and community goals, as well as punishing non-contributors," is an ideal framework for studying the altruistic punishment behavior (Gintis, 2000 171). The experimental observation of a particular behavior is important because it empirically confirms that humans actually behave that way. However, there must also be an explanation for what motivates this behavior in rational individuals. I will discuss three prevailing theories of proximate motivations for altruistic punishment.

Moreover, the behavior, and the motivation for the behavior must also have an origin. Here I will explore arguments in the scientific literature for how altruistic punishment could have evolved in human societies. The discussion of the evolution of altruistic punishment is the second reason why I focus on public goods games. Bowles and Gintis argue that "many human interactions in the relevant evolutionary context took the form of *n*-person public goods games—food sharing and other co-insurance, as well as common defense—rather than dyadic interactions" (Bowles and Gintis 26). Public goods games accurately model the type of cooperative endeavor that early human communities faced. I will begin with rejected evolutionary explanations and then turn to a discussion of group-selection research, whose models make sense of human strong reciprocity. After explaining several group-selection theories of the evolution of punishment and cooperation, I will also discuss criticisms and proposed alternatives. I will end with a discussion that explains why I find some models and data convincing and others in need of future research and analysis.

I believe that altruistic punishment provides a practical explanation for cooperation in modern human societies. However, modeling the course of the evolution of human behavior is speculative and dependent on variable dynamics. Sustaining cooperation in vast social groups is

as complex as the number of individual participants who decide whether to behave pro-socially. We will likely never have a definitive explanation of the origins and evolution of human cooperation. Yet altruistic punishment may form an integral part of that trajectory.

Laboratory Experiments: Evidence of Strongly Reciprocal Behavior

Game Theorists predict players will not contribute to a public good *"if it is common knowledge that all subjects are rational and selfish money-maximizers*[ii]" (Fehr, Fischbacher and Gächter 13). Non-contribution is the dominant strategy because a public good "can be consumed by every group member regardless of the member's contribution to the good. Therefore, each member has an incentive to free-ride on the contributions of others" (Fehr and Fischbacher, 2003 786). Moreover, a player has no incentive to cooperate because "any form of cooperation causes a reduction in the material payoff to the cooperating subject" (Fehr, Fischbacher and Gächter 13). Thus, no player should ever contribute to a public good.

Punishment could create a material incentive that would encourage players to contribute. However, backwards induction confirms that a selfish player would not punish others in a one-shot public goods game. If a game ends in 10 rounds, then a player's "best choice at the punishment stage in period ten is not to punish at all because punishment is costly" (Fehr, Fischbacher and Gächter 13). Other players will realize that punishment is a suboptimal strategy, and will neither fear punishment nor elect to contribute. Accordingly, "the presence of the punishment stage does not change the behavioral incentives at the investment stage of period ten. Therefore, in the *punishment condition* also nobody will invest in period ten" (Fehr, Fischbacher and Gächter 13). Using this logic, we can rollback through all stages of the game "until period one so that full defection and no punishment is predicted to occur for all ten periods of the punishment treatment" (Fehr, Fischbacher and Gächter 14). Thus, the dominant strategy for a

utility maximizing player in a public goods game with or without punishment is to free-ride on the contributions of others.

Yet test subjects do not behave this way in laboratory experiments. I will begin with experiments in which there was no explicit punishment opportunity. Fehr and Schmidt (1999) did a comprehensive study of public goods games. In this "meta-study of 12 public goods experiments" they "found that in the early rounds, average and median contribution levels ranged from 40 to 60% of the endowment, but in the final period 73% of all individuals ($N = 1042$) contributed nothing, and many of the remaining players contributed close to zero" (Gintis, 2000 171). Contribution to the public goods game is "rarely stable and deteriorates to rather low levels if the game is played repeatedly (and anonymously) for ten rounds" (Fehr and Fischbacher, 2003 786). This deterioration of cooperation indicates that "*full free-riding emerges as the focal individual action*" in the absence of punishment (Fehr and Gächter, 2000 986).

Game theorists might be perplexed that test subjects contribute in the first period and continue to contribute in later stages of the public goods game; they predict universal defection from stage one. Individuals may contribute in the initial condition because they do not understand the game. Andreoni (1995) surveyed test subjects after participating in a public goods game and found that "Confusion is by far the dominant motive in round 1 of the experiment, accounting for 81 percent of all cooperation" yet by round ten confusion was reduced "to a mere 13.6 percent in round 10" (Andreoni 897). I think a better explanation is that people are not purely selfish; they may be willing to cooperate by contributing to the public good because they expect that other people will also contribute, thus improving the group's collective payoff. This strategy is called reciprocal altruism, and this type of individual "increases his contribution levels

in response to expected increases in the average contribution of other group members" (Fehr and Fischbacher 786). Reciprocal altruism explains contributions; selfishness cannot.

Researchers have also tried to explain why subjects who initially cooperated began to defect in later rounds. Some claim that the presence of defectors will cause cooperators to stop cooperating. Initial contributions followed by declining participation "might be predicted by a reciprocal altruism model, since the chance to reciprocate declines as the end of the experiment approaches. (Gintis, 2000 171). Without the opportunity to punish, "the selfish types induce the reciprocal types to behave noncooperativley, too" (Fehr, Fischbacher and Gächter 15). In some interpretations of the public goods game, a "single selfish player is capable of inducing all other players to contribute nothing to the public good, although the others may care a lot about equity" (Fehr and Schmidt 819). These researchers believe that altruistic test subjects forfeit their beliefs when in non-cooperative situations.

I think that there is a better explanation. Andreoni (1995) found that test subjects who had contributed to the public good claimed that they "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" (Gintis, 2000 171). This response indicates that subjects are not reciprocal altruists: they did not reduce their contributions because of diminished opportunities for reciprocity. Instead, "Noncooperation is the only way in which the reciprocal types can at least implicitly punish the defectors in their groups" (Fehr, Fischbacher and Gächter 15-6). The test subjects' natural reaction was a desire to punish others for not participating in the public goods game, a reaction not predicted by theories of reciprocal altruism.

The finding that test subjects' behavior did not coincide with either purely self-interested or purely-altruistic models was further confirmed when subjects were given the explicit

opportunity to punish. The experimental results are drastically different from the game-theory prediction of universal defection: "a strikingly large fraction of roughly 80 percent cooperates *fully* in the game with punishment" (Fehr and Schmidt 838). Fehr and Gächter (2000, 2002) examined this phenomenon extensively, allowing players to punish others in both a "Stranger" and "Partner" treatment of a public goods game. In the Partner treatment, group composition remains stable for ten periods of a public goods game whereas in the Stranger treatment, group composition "randomly changes from period to period" (Fehr and Gächter, 2000 981). In the results of both the Stranger and Partner treatment, "*toward the end there is a relative payoff gain in both treatments*" and "*a subject is more heavily punished the more his or her contribution falls below the average contributions of other group members*" (Fehr and Gächter, 2000 993, 990). Moreover, free-riders' payoffs are reduced dramatically from the no-punishment condition: by 24 percent in the Stranger treatment and by 16 percent in the Partner treatment and not less only because contributors "also contribute more in the punishment condition" (Fehr and Gächter, 2000 992). What is most similar about the Partner and Stranger treatments is that "Spontaneous and uncoordinated punishment activities give rise to heavy punishment of free-riders" (Fehr and Gächter, 2000 993). It is surprising that punishment occurs in the Stranger treatment, "although it is costly and provides no future private benefits for the punishers" but remarkable that "the strength of the punishment is almost as high in the Stranger design as in the Partner design" (Fehr and Gächter, 2000 993; Fehr, Fischbacher and Gächter 15). In conclusion, Fehr and Gächter write that punishment opportunities "completely remove the drawing power of the equilibrium with complete free-riding" and sustain a cooperative equilibrium (Fehr and Gächter (2000) 985). While "the presence of punishment opportunities eventually leads to pecuniary efficiency gains," the presence of punishing strategies alone does not ensure cooperation (Fehr

and Gächter, 2000 993). Instead, "It is not only the punishment opportunity (that is, the non-executed punishment threat but also the actual punishment that raised cooperation levels" (Fehr and Gächter, 2002 138). Punishers cannot simply threaten to punish defectors; they must actually incur the cost of punishment in order to induce defectors to contribute to the public good.

The test subjects' behavior described by numerous experiments with public goods games confirms the same outcome: people contribute to the public good in games with and without punishment, but widespread cooperation can only be sustained in the punishment condition. Punishers were obviously not purely selfish, because they punished others, both directly and indirectly, at a material cost to themselves. Yet they were not purely altruistic either: the use punishment "is clearly inconsistent with models of pure altruism" because "an altruistic person never uses a costly option to reduce other subjects' payoffs" (Fehr and Gächter, 2000 993). Instead, this punishing behavior has been described as a unique behavior: altruistic punishment or strong reciprocity. Strong reciprocity or altruistic punishment occurs when "people tend to behave prosocially and punish antisocial behavior, at a cost to themselves, even when the probability of future interactions is extremely low, or zero" (Gintis, 2000 177). In comparison, reciprocal altruism is weak reciprocity, because the altruistic behavior is contingent on the actions of others. Non-altruistic punishers follow a "hypocritical strategy" where they do not contribute to the public good "while urging others to cooperate through participation in the sanctioning system" (Heckathorn 80). Moreover, reciprocal altruists "reward and punish only if this in their long-term self-interest" (Fehr and Fischbacher, 2003 785). Conversely, "Strong reciprocators bear the cost of rewarding or punishing even if they gain no individual economic benefit whatsoever from their acts" (Fehr and Fischbacher, 2003 785). Strong reciprocity and altruistic punishment are synonyms[iii]: they imply that an individual contributes to the public

good and punishes others at a personal cost in order to sustain a cooperative norm that redounds to the benefit of her social group. I will use both terms, interchangeably, simply depending on what term the researcher used in her own analysis of the behavior.

One potential criticism of these public goods experiments is that the punishment strategy was not actually altruistic. Perhaps the subjects were incentivized to behave in this way for their own self interest or financial gain. Fehr and Gächter (2000, 20002) argue that their experimental design completely prevented the punishment strategy from being anything other than altruistic. If test subjects could develop "an *individual* reputation[iv] " as a cooperator or a defector, then "there were material incentives for cooperation and for punishment," namely to cooperate even if one wanted to defect because punishment could be targeted at one's individual identity (Fehr and Gächter, 2000 981). To remove such a material incentive, "we eliminated all possibilities for individual reputation formation and implemented treatment conditions with an *ex ante known* finite horizon," for example by using computers as an interface between subjects and listing contributions randomly, without any reference to which individual test subject made a particular contribution (Fehr and Gächter, 2000 981). Moreover, in the Stranger treatment, the test subjects changed partners; "group composition changed from period to period such that no subject ever met another subject more than once" (Fehr and Gächter, 2002 137). This ensures altruistic punishment in two ways. While "punishment may well benefit the future group members of a punished subject," the punisher receives no benefit "because the punishing subject never meets the same subjects again" (Fehr and Gächter, 2002 137; 138-9). Second, purely selfish subjects will not contribute: punishment is individually costly and they cannot be identified as defectors when group membership is not static. Thus, "The selfish motives associated with theories of indirect reciprocity or costly signaling cannot explain cooperation and punishment in this

environment" (Fehr and Gächter, 2002 137). Accordingly, altruistic punishment was definitively observed in these public goods games. This behavior requires explanation.

Proximate Explanations for Strongly Reciprocal Behavior

Altruistic punishment implies that "individuals have proximate motives beyond their economic self-interest—their subjective evaluations of economic payoffs differ from the economic payoffs" (Fehr and Fischbacher, 2003 788). I will consider three proximate explanations for strongly reciprocal behavior. The first explanation for why researchers see strong reciprocity in laboratory experiments is that defection causes negative emotions in contributors. Fehr and Gächter (2002) argue: "Free riding may cause strong negative emotions among the cooperators and these emotions, in turn, may trigger their willingness to punish the free riders" (Fehr and Gächter, 2002 139). Their results indicate that negative emotions trigger altruistic punishment for three reasons. First, "if negative emotions trigger punishment, most punishment acts will be expected to be executed by above-average contributors" (Fehr and Gächter, 2002 139). Their results confirmed that altruistic punishers contributed to the public good and that defection lowers the value of the mean contribution to below what either punishers or contributors contribute. Second, "The intensity of negative emotions towards a free rider varies with the deviation from the others' average contribution;" their results confirmed that "punishment increased with the deviation of the free rider from the average investment of the other members" (Fehr and Gächter, 2002 139). Finally, "if negative emotions cause punishment, the punishment threat is rendered immediately credible because most people are well aware that they trigger strong negative emotions when they free ride" (Fehr and Gächter, 2002 139). When surveyed after the experiments, defecting test subjects "seemed to have had a clear understanding of why they were punished and how they should respond to the punishment," "immediately

changing" from defection to contribution (Fehr and Gächter, 2000 992). Moreover, low contributors "expected a higher intensity of negative emotions," likely because they "experience more sanctions in the punishment condition" (Fehr and Gächter, 2002 139). In sum, altruistic punishers become angry when others fail to contribute and severely punish those who contribute the least. Moreover, selfish test subjects expect to be punished harshly when they make low contributions, and most harshly when they defect. Hence, "These observations are consistent with the view that emotions are an important proximate factor behind altruistic punishment" (Fehr and Gächter, 2002 139).

Fehr and Schmidt argue that people do not have strong emotions simply when others defect. Instead, people have negative emotions when others defect because defection creates an unfair outcome. Fairness is defined as "self-centered inequity aversion," which means that individuals want to avoid "inequitable outcomes" and "are willing to give up some material payoff to move in the direction of more equitable outcomes" (Fehr and Schmidt 819). "Self-centered" means that individuals only care about inequities in "their own material payoff relative to the payoff of others" (Fehr and Schmidt 819). Free-riding "generates a material payoff disadvantage relative to those who cooperate;" hence cooperators, who become "sufficiently upset by the inequality to their disadvantage," "are willing to punish the defectors even though this is costly to themselves" (Fehr and Schmit 840). Moreover, "the more these enforcers care about disadvantageous inequality, the more they are prepared to punish defectors" (Fehr and Schmidt 842). Cooperation is sustained when people care a lot about inequity because individuals who are inclined to defect will cooperate when punishment has credibility (Fehr and Schmidt 840). Fehr and Schmidt conclude that "psychological evidence on social comparison

and loss aversion" justifies their thesis that altruistic punishment can be explained by a concern for "equitable outcomes" (Fehr and Schmidt 866).

However, these theories of motivation still do not arrive at the root of why people behave this way. Neurology may provide an answer. Because altruistic punishment is "an action based on deliberation and intent, humans have to be motivated to punish. The typical proximate mechanism for inducing motivated action is that people derive satisfaction from the action" (de Quervain et al. 1254). People "seem to feel bad if they observe that norm violations are not punished, and they seem to feel relief and satisfaction if justice is established" (de Quervain et al. 1254). If an individual is satisfied when justice is established, and is willing to punish because she "anticipates deriving satisfaction from punishing, we should observe activation predominantly in those reward-related brain areas that are associated with goal-directed behavior" (de Quervain et al. 1254). Study findings confirm this hypothesis. Experimental results indicate that altruistic punishment caused caudate activation: this is the area of the brain "implicated in making decisions or taking actions that are motivated by anticipated rewards" (de Quervain et al. 1258). Moreover, subjects' neurological response to punishment was observed under two different conditions—costless and costly, or altruistic, punishment. Subjects that exhibit "higher caudate activation at the maximal level of punishment if punishment is costless for them also spend more resources on punishment if punishment becomes costly" (de Quervain et al. 1258). This means that "high caudate activation seems to be responsible for a high willingness to punish" and that "caudate activation reflects the anticipated satisfaction from punishing defectors" (de Quervain et al. 1258). When strong reciprocators care about fairness, such as the inequity aversion suggested by Fehr and Schmidt, they anticipate deriving satisfaction from punishing others and are willing to punish even at a material cost to themselves.

Because altruistic punishment activates the areas of the brain that anticipate rewards, "humans may have physically or developmentally evolved this behavior" (Fowler 7047). I will now explore theories of how this behavior could have evolved.

Contested Explanations for the Evolution of Altruistic Punishment and Cooperation

First, I will clarify the meaning of altruistic behavior. Altruism has distinct meanings depending on the academic background of the researcher. For psychologists, whether an act is altruistic is contingent on the intentions of the actor; this definition "requires that the act be driven by an altruistic motive that is not based on hedonic reward" (de Quervain et al. 1257). Biologists have a different view. An act is altruistic "if it is costly for the actor and confers benefits on other individuals. It is completely irrelevant for this definition whether the act is motivated by the desire to confer benefits on others, because altruism is solely defined in terms of the consequences of behavior" (de Quervain et al. 1257). Because altruistic punishment is costly to the individual and benefits the group by inducing "the punished individual to defect less in future interactions with others," "the punishment of defectors is an altruistic act in the biological sense" (de Quervain et al. 1257). However, "our results suggest that it is not an altruistic act in the psychological sense" (de Quervain et al. 1257). Thus, the biological definition of altruism is germane to discussions of the evolution of altruistic punishment. It is important to understand that the selecting forcing that drove the evolution of altruistic punishment had nothing to do with the motivations of the actors engaged in that evolutionary process.

Researchers argue that the evolution of altruistic punishment can support observations of cooperation amongst humans better than any other explanation for human cooperation. Aside from research on strong reciprocity, "human cooperation has mainly been explained in terms of kin selection, reciprocal altruism, indirect reciprocity and costly signaling," which sustain

cooperation through "mechanisms other than altruistic punishment" (Fehr and Gächter, 2002 139). Researchers maintain that "strong reciprocity cannot be rationalized as an adaptive trait" by these "major prevailing evolutionary theories" (Fehr, Fischbacher and Gächter 20). I will examine the criticisms of each of these theories and then turn to discussions of the evolution of strong reciprocity.

Kin-selection theory cannot account for strongly reciprocal behavior in modern societies. People may cooperate in small groups, such as family units, because they care about the evolutionary fitness of their direct kin. Thus, they cooperate precisely because they are related; this is a likely reason for cooperation in early human evolution when social groups were constrained to genetic relatives. Yet, "As group size rises above 10, to 100 or 1000, cooperation is virtually impossible to evolve or maintain with only reciprocity and kinship" because the likelihood that the individuals are related decreases as group size increases (Henrich and Boyd 79).

Next, researchers think that theories of indirect reciprocity confuse cause with effect. Bowles and Gintis (2004) "doubt that indirect reciprocity can be sustained in a population of self-interested agents" (Bowles and Gintis 26). Instead, "Indirect reciprocity is more likely promoted, as in our model, by strong reciprocators who reward prosocial behavior and punish anti-social behavior even when this behavior reduces within-group fitness" (Bowles and Gintis 26). Therefore, they suggest that strong reciprocity induces both cooperation and indirect reciprocity. Indirect reciprocity alone is insufficient to maintain a cooperative equilibrium.

Likewise, costly signaling theories consider altruistic punishment a costly signal without acknowledging that altruistic punishment is an independently sufficient means to cooperation. In one signaling model, punishment "is the benefit to others that signals high-quality. Our model

easily allows such punishment or enforcement to serve as the costly signal, and hence to be maintained when the conditions for evolutionary stability specified in the model are met" (Gintis et al. 116). It may be true that altruistic punishment can be defined as a costly signal. However, strong reciprocity can also evolve through other dynamics and sustain cooperation without being modeled as a costly signal. Even if costly signaling is to be accepted, "the role that costly signaling might play in enforcement of prosocial behavior is as yet untested, but deserves further investigation" (Gintis et al. 116). Proponents of the evolution of altruistic punishment claim that it is directly responsible for the enforcement of prosocial behavior. The link between costly signaling and punishment deserves further attention before a causal relationship between signaling, punishment and cooperation can be established.

The theory that has achieved the most credit for sustaining cooperation is reciprocal altruism. In fact, "many behavioral scientists believe that reciprocal altruism is sufficient to explain human sociality" (Gintis 177). The reciprocal altruist—or conditional cooperator—cooperates in the public goods game but only insofar as other players also cooperate. For a reciprocal altruist, "the only evolutionarily stable strategy in the *n*-person public goods game is to cooperate as long as all others cooperate and to defect otherwise" (Gintis et al. 164). While this equilibrium may be sustained in small groups, "the basin of attraction of this equilibrium becomes very small as group size rises, so the formation of groups with a sufficient number of conditional cooperators is very unlikely" (Gintis et al. 164). Moreover, this equilibrium "can be disrupted by idiosyncratic play, imperfect information about the play of others, or other stochastic events" and is "a 'knife-edge' that collapses if just one member deviates" (Gintis et al. 164). The reciprocal altruism equilibrium, albeit unlikely to arise or to be sustained, does create cooperation. However, it is extremely inefficient: a reciprocal altruist "withdraws cooperation in

retaliation for the defection of a single group member," thus inflicting "punishment on all members, defectors and cooperators alike" (Gintis et al. 164). The strategy of reciprocal altruism punishes those who contribute to the public good; this punishment is ultimately inimical to the end of fostering cooperative behavior.

A second argument against reciprocal altruism arises from a historical evaluation of human evolution. This argument is integral in the Gintis (2000) model discussed below. In essence, reciprocal altruists have no incentive to contribute to a public good when their "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" (Bowles and Gintis 26). Thirdly, researchers assert that uniquely human behavior makes reciprocal altruism an inadequate explanation. Empirically, "the contemporary study of human behavior has a documented a large class of prosocial behaviors inexplicable in terms of reciprocal altruism" (Bowles and Gintis 26). Researchers argue that "the evolutionary success of our species and the moral sentiments that have led people to value freedom, equality, and representative government are predicated upon strong reciprocity and related motivations that go beyond inclusive fitness and reciprocal altruism" (Gintis et al. 144). Evolutionary models that explain strong reciprocity, including group selection, multilevel selection of group and culture, and gene and culture coevolution, may account for the origins of punishing behavior that theories of reciprocal altruism cannot explain.

The Evolution of Altruistic Punishment in Humans: Recent Theories of Group Selection, Multi-level Selection, and Gene and Culture Co-Evolution

The most recent body of research about the origins of cooperation in public goods games postulates the evolution of altruistic punishment or strong reciprocity through group selection. These models have two integral features in common. Boyd, Gintis, Bowles and Richerson (2003)

concisely illustrate both features. The first feature is that the mode of punishment in these models is altruistic: Punishers incur costs for punishing defectors. Where *k*, *x*, and *y* denote cost of punishment, frequency of contributors and frequency of defectors, respectively, "Punishers suffer a fitness disadvantage of *k*(1—*x*—y) compared with nonpunishing contributors" (Boyd et al. 3531). Thus, a punisher's fitness advantage is reduced by the cost of punishment multiplied by the frequency of defectors in the population.

The second feature is that contributors have a fitness advantage to punishers when they cooperate in the public goods game but do not punish defectors. This behavior is called second order free-riding. The more punishment that altruistic punishers must dole out "increases the payoff advantage of second order free riders compared with altruistic punishers," making the contributors substantively more fit (Boyd et al. 3534). Therefore, second-order free-riding creates a payoff asymmetry between altruistic punishers and non-punishing contributors, even though both groups contribute equally to the public goods game. However, contributors will have a higher payoff than defectors "if punishers are sufficiently common that the cost of being punished exceeds the cost of cooperating (*py>c)*" (Boyd et al. 3531). The presence of punishers makes defecting a suboptimal strategy; players are incentivized to cooperate in the public goods game. Thus, "the payoff disadvantage of punishers relative to contributors approaches zero as defectors become rare because there is no need for punishment" (Boyd et al. 3531). In the absence of this payoff asymmetry, contributors and altruistic punishers are equally fit and the cooperative equilibrium yields the highest payoff to both types.

Group selection models must take this payoff asymmetry into account. Because punishment is individually costly, "within-group selection creates evolutionary pressures against strong reciprocity" (Fehr, Fischbacher and Gächter 5). However, the presence of altruistic

punishers or strong reciprocators allows groups to reach the most cooperative equilibrium. When cooperative behaviors make the group more fit, cooperative groups will survive. Therefore, "between-group selection favors strong reciprocity because groups with disproportionately many strong reciprocators are better able to survive" (Fehr, Fischbacher and Gächter 5). Other groups will imitate whatever strategy the successful groups have adopted; because altruistic punishment is the most fitness enhancing strategy, strongly reciprocal behavior will proliferate. However, "the consequence of these two evolutionary forces is that in equilibrium strong reciprocators and purely selfish humans coexist" (Fehr, Fischbacher and Gächter 5). Group selection theory thus explains how strong reciprocity could sustain a cooperative equilibrium but does not claim that all individuals will be altruistic punishers.

Boyd, Gintis, Bowles and Richerson (2003) use a "modest" group selection model to show that altruistic punishment can sustain a cooperative equilibrium in large groups when altruistic cooperation alone cannot. Group selection "acts to favor individually costly, group beneficial behaviors," such as altruistic punishment (Boyd et al. 3534). First, altruistic cooperation has a strict evolutionary disadvantage relative to altruistic punishment. This is due to a payoff asymmetry. For altruistic cooperators, the payoff disadvantage "relative to defectors is independent of the frequency of defectors in the population" (Boyd et al. 3531). However, the payoff disadvantage for altruistic punishers "declines as defectors become rare because acts of punishment become very infrequent. Thus, when altruistic punishers are common, individual level selection operating against them is weak" (Boyd et al. 3531). Altruistic cooperation and punishment therefore sustain very different levels of cooperation. Without punishment, "group selection can support high frequencies of cooperative behavior only if groups are quite small"

(Boyd et al. 3533). Whereas altruistic cooperation cannot survive in large groups because of prohibitive costs, altruistic punishment can be an evolutionarily stable strategy.

Second, their model allows groups to imitate the most successful strategy; this "payoff biased imitation strategy maintains variation among groups in the frequency of cooperation" (Boyd et al. 3534). Groups that contain punishers "will tend to exhibit a greater frequency of cooperative behaviors (by both contributors and punishers)" and thus "the frequency of punishers and cooperative behaviors will be positively correlated across groups" (Boyd et al. 3531). As defection decreases, the punishers' payoff disadvantage relative to contributors also decreases, and "as a result, variation in the frequency of punishers is eroded slowly" (Boyd et al. 3534). Moreover, "in groups in which punishers are common, defectors achieve a low payoff and are unlikely to be imitated" (Boyd et al. 3534). Subsequently, altruistic punishers are more fit than defectors, and the imitation of "punishment will increase as a 'correlated response' to group selection that favors more cooperative groups" (Boyd et al. 3531). To the extent that cooperative groups are more evolutionarily fit, punishing behavior will proliferate because it best sustains cooperation.

Herbert Gintis (2000) focuses on group extinction in early human evolution to demonstrate that strong reciprocity can sustain punishment where other theories of altruism cannot. Previous theories of the emergence of altruism, such as reciprocal altruism, "tended to argue the plausibility of altruism in general, rather than isolating particular human traits that might have emerged from a group selection process" (Gintis, 2000 169). Strong reciprocity is a possible group-selection trait, "an empirically identifiable form of prosocial behavior in humans that probably has a significant genetic component" (Gintis, 2000 169). In his analysis of Fehr and Gächter's public good experiments, he suggests that test subjects are motivated by "the

personal desire to punish free riders (the stranger treatment), but even more strongly motivated when there is an identifiable group, to which they belong, whose cooperative effort is impaired by free riding (the partner treatment)" (Gintis, 2000 172). Strong reciprocity better sustains cooperation "the more coherent and permanent the group in question;" hence, this behavior been described as "prosocial" (Gintis, 2000 172). The prosociality of strong reciprocity is integral to the evolutionary explanation of cooperation in human society. In terms of evolutionary dynamics, if strong reciprocity is a trait that evolved through group selection, then "it must be a considerable benefit to a group to have strong reciprocators, and the group benefits must outweigh the individuals' costs" (Gintis 173). Gintis proves that in the context of group extinctions, strong reciprocity could have evolved through group selection.

Gintis' unique idea is that remaining a member of a group has a greater utility for any individual than free-riding in the public goods game. If an individual who fails to contribute is punished through ostracization, then cooperation is attainable (Gintis, 2000 170). Without altruistic punishment, "*if groups disband with high probability, then cooperation among self-interested agents cannot be sustained*" (Gintis, 2000 170). Human groups were likely threatened by a variety of forces, including internal strife or environmental threats, like drought, that could result in the disintegration of the group. Gintis provides empirical evidence, such as "flattened mortality profiles of pre-historic skeletal populations," that suggests "periodic social crises are not implausible" (Gintis, 2000 170) A self-interested individual will not cooperate in the face of these threats: not only does "the threat of ostracism" lose its disutility but also "future gains from cooperation become very uncertain" when "the probability that the group will dissolve becomes high" (Gintis et al. 163). When a human group is threatened with extinction, "reciprocal altruism will fail to motivate self-interested individuals in such periods, thus exacerbating the threat and

increasing the likelihood of group extinction" (Gintis, 2000 177). Reciprocal altruism cannot sustain cooperation under such conditions: "*precisely when a group is most in need of prosocial behavior, cooperation based on reciprocal altruism will collapse*, since the discount factor then falls to a level rendering defection an optimal behavior for self-interested agents" (Gintis 172).

However, strong reciprocity can sustain cooperation even when groups are likely to disband. A strong reciprocator behaves differently than a purely self-interested individual, because she "cooperates and punishes non-cooperators without considering the value of δ, i.e. even when the probability of future interactions is low (Gintis, 2000 170). Unlike the self-interested actor who will stop cooperating when he fears that the group will separate, the strong reciprocator continues to cooperate in the public goods game and punish defectors. Strong reciprocators can also generate cooperation from the non-punishers: "If the fraction of strong reciprocators is sufficiently high, even self-interested agents can be induced to cooperate in such situations, thus lowering the probability of group extinction" (Gintis, 2000 178). Moreover, the fraction of strong reciprocators that is required to sustain cooperation can be quite low, for example if the probability of facing a threat is low and surviving is high (Gintis, 2000 174). When there are a sufficient number of strong reciprocators to enable cooperation in a group, that group "will then outcompete other self interested groups, and the fraction of strong reciprocators will grow. This will continue until an equilibrium fraction of strong reciprocators is attained" (Gintis et al. 163). Therefore, strong reciprocity facilitates group survival by sustaining cooperation and "might even have an evolutionary advantage in situations where groups are frequently threatened" (Gintis, 2000 172-3)

Gintis continued his work on the evolution of strong reciprocity with a multi-level selection theory in collaboration with Samuel Bowles. Bowles and Gintis (2004) de-emphasize

extinctions, which played an important role in Gintis' earlier work, instead pursuing other features of early human social history. Their theory is that in early human societies "punishment takes the form of ostracism or shunning, and those punished in this manner suffer fitness costs" (Bowles and Gintis 17-8). This model is based on hypothetical socio-biological human history, "on the structure of interaction among members of the mobile hunter-gatherer bands in the late Pleistocene" (Bowles and Gintis 18). They select this period for a variety of reasons, including sufficiently large group size to allow for free-riding and not allow for kinship explanations (Bowles and Gintis 18). Perhaps most importantly, they explain that ostracism in this stage of human development could serve as a legitimate punishing strategy. Bowles and Gintis explain that the cost of ostracism is entirely contingent on the stage of human sociological evolution: "we treat the cost of being ostracized as endogenously determined by the amount of punishment and the evolving demographic structure of the populations" (Bowles and Gintis 18). Further, ostracism is most appropriate to this model because this punishment "reflects a central aspect of hunter-gatherer life: since individuals can often leave the group to avoid punishment, the cost of being ostracized is among the more serious penalties that can be levied upon an individual group member" (Bowles and Gintis 18). Ostracism as punishment makes sense, then, in the absence of alternatives. Bowles and Gintis explain that at this stage of human development, property was communal and shelter was limited; therefore, property loss and confinement were not potential punishment strategies (Bowles and Gintis 18).

Moreover, Bowles and Gintis contend that "strong reciprocity is exhibited in such collective situations as group food-sharing and defense," the types of cooperative activities of hunter-gatherer groups (Bowles and Gintis 26). Previous models of the evolution of cooperation that focused on interactions between two individuals do not appropriately capture the type of

cooperation that occurred in our evolutionary history. Instead, they model "*n*-agent groups (where *n* is on the order of ten to 100) in a series of production periods that are effectively one-shot, since the only inter-period influences are those involving the biological and cultural reproduction of new agents" (Bowles and Gintis 26). Thus, they eliminate considerations of confounding factors like reputation by focusing exclusively on group selection and fitness.

An individual's type, as a reciprocator, cooperator or selfish, is a genetic predisposition. An offspring takes on her parents' type "with probability 1-ε, and with probability ε/2, an offspring takes on each of the other two types. We call ε the *rate of mutation*" (Bowles and Gintis 19). Where "members of a group benefit from mutual adherence to a norm," a strong reciprocator will "obey the norm and punish its violators, even when this behavior incurs fitness costs" (Bowles and Gintis 18). Thus, when a norm of cooperation has been established, strong reciprocators will contribute and punish to sustain that norm. They find that strong reciprocity could emerge "after as few as 500 periods" because "it does not take that many periods before at least one group will have enough Reciprocators to implement a high level of cooperation" (Bowles and Gintis 25). Strong reciprocity spreads when groups reproduce, "and as a result it seeds other groups by migration and repopulates the sites of disbanded groups" (Bowles and Gintis 25). Thus, strong reciprocity can spread rather rapidly "for the simple reason that in order to proliferate the behavior need only become common in a single group" (Bowles and Gintis 25). In sum, this model can "capture the environments that may have supported high levels of cooperation among our ancestors living in mobile foraging bands during the late Pleistocene" (Bowles and Gintis 27). As a socio-biological history, doubts remain whether strong reciprocity really evolved in this way, "but our simulations suggest that it could have" (Bowles and Gintis 27).

Henrich and Boyd (2001) contend that altruistic punishment and cooperation could have evolved through cultural group selection. Their thesis is that "the evolution of cooperation and punishment are plausibly a side effect of a tendency to adopt common behaviors during enculturation" (Henrich and Boyd 80). Humans undergo a process of socialization whereby they learn what behaviors to adopt in order to enhance their fitness. Further, "humans do not simply copy their parents, nor do they copy other individuals" but instead use "social learning rules" to select the best strategy to imitate (Henrich and Boyd 80). These rules or "short-cuts" include "pay-off biased transmission," or imitate-the-successful, and "conformist transmission," or imitate-the-majority (Henrich and Boyd 80). Henrich and Boyd focus on conformist transmission. Because "not-cooperating leads to higher payoffs than cooperating," pay-off biased transmission cannot explain cooperation in the absence of punishment (Henrich and Boyd 81). However, the presence of altruistic punishers makes defection a suboptimal strategy. If altruistic punishment maximizes payoffs, individuals will imitate this strategy. "Individuals preferentially adopt common behaviors" when they follow a conformist transmission shortcut, "which acts to increase the frequency of the most common behavior in the population" (Henrich and Boyd 81). Thus, social learning shortcuts create a self-enforcing equilibrium: pay-off biased transmission causes individuals to imitate altruistic punishment and conformist transmission spreads the behavior once it is common. By allowing individuals to punish in at least two periods, "a relatively weak conformist tendency can stabilize punishment and therefore cooperation" (Henrich and Boyd 86).

Yet in terms of genetic evolution, "the stabilization of punishment is, from the gene's point of view, a maladaptive side-effect of conformist transmission" (Henrich and Boyd 81). If "there were genetic variability in the strength of conformist transmission" and "cooperative

dilemmas were the only problem humans faced," humans would not imitate the most prevalent strategy and altruistic punishment could not evolve (Henrich and Boyd 81). However, social learning short-cuts such as conformist transmission have evolved precisely because they allow humans "to efficiently acquire adaptive behaviors over a wide range of behavioral domains and environmental circumstances," not just in cooperative endeavors (Henrich and Boyd 81). As long as "distinguishing cooperative dilemmas from other kinds of problems is difficult, costly or error prone," humans have an incentive to imitate the majority (Henrich and Boyd 82). Because "it is difficult to imagine a cognitive mechanism capable of distinguishing cooperative circumstances" from other human behaviors, we can assume that conformist transmission would allow for the stabilization of punishment and cooperation (Henrich and Boyd 82).

Conformist transmission does not yield a unique cooperative equilibrium. This model allows for a second equilibrium of "non-cooperation and non-punishment;" yet Henrich and Boyd contend that populations will stabilize at the former equilibrium because of cultural group selection (Henrich and Boyd 86). Their model suggests that "cultural evolutionary processes will cause groups to exist at different behavioral equilibria;" this means that groups exist with varying degrees of punishers, cooperators and non-cooperators and accordingly different payoffs (Henrich and Boyd 86). Cultural group selection requires that these variances between groups exist and also that the social learning shortcuts be "strong enough to maintain stable cooperation in the face of migration between groups" (Henrich and Boyd 86). Cultural group selection allows "prosocial behavior," such as altruistic punishment, to spread in several ways: cooperative groups will outcompete non-cooperative groups because they will have more public goods, such as armies, and greater reproduction rates because all group members have more capital (Henrich and Boyd 86). If people follow a pay-off biased transmission model, then they will imitate

cooperators. If individuals can imitate people in any group, "people from cooperative populations will be preferentially imitated by individuals in non-cooperative populations because the average payoff to individuals from cooperative populations is much higher than the average payoff of individuals in non-cooperative populations" (Henrich and Boyd 87). This means that pro-social behaviors can spread from "a single group (at a group-beneficial equilibrium) through a meta-population of other groups, which were previously stuck at a more individualistic equilibrium" (Henrich and Boyd 87).

If the meta-population of groups achieves the cooperative equilibrium, then defectors are at a severe payoff disadvantage. Accordingly, "it is plausible that natural selection acting on genetic variation will favor genes that cause people to cooperate and punish—because such genes decrease an individual's chance of suffering costly punishment" (Henrich and Boyd 87). This genetic evolution could occur in a variety of ways. The mechanism is not as important as the consequence: "As pro-social genes spread among groups with different stable cooperative domains, individuals with such genes would be more likely to mistakenly cooperate in non-cooperative cultural domains" (Henrich and Boyd 88). They argue that human groups may not cooperate in all activities; "cooperation may not be a dispositional trait of individuals, but rather a specific behavior or value tied only to certain cultural domains" (Henrich and Boyd 88). For example, a cultural group may cooperate in hunting but not in cooking. A migrant individual with prosocial genes who mistakenly cooperates in cooking will not be punished, though she may suffer a payoff reduction for wasting her time; overall, her prosocial behavior in this activity "will be comparatively neutral in non-cooperative populations" (Henrich and Boyd 88). However, "prosocial genes will be favored in a wide range of circumstances in cooperative populations" and defectors, or those who lack prosocial genes, will be at a strict evolutionary

advantage when they are punished for not cooperating (Henrich and Boyd 88). Thus, Henrich and Boyd demonstrate not only that punishment can sustain cooperation in a group, but also that cooperative behaviors can migrate across groups, and may even influence our genetic makeup.

In his more recent work, Gintis (2003) continues along this trajectory, examining the theory of gene-culture co-evolution. He explains the evolution of pro-social norms through both genetic and cultural selection processes. Gintis develops a "Multi-level gene-culture coevolutionary model to elucidate the process whereby altruistic internal norms will tend to drive out norms that are both socially harmful and individually fitness-reducing" (Gintis, 2003 408). He explains that an individual internalizes a norm through socializing forces, such as parenting (Gintis, 2003 407). Internal norms have unconditional value; an individual will behave this way "because they value this behavior for its own sake, in addition to, or despite, the effects the behavior has on personal fitness and/or perceived well-being" (Gintis, 2003 408). Gintis continues that an instrumental norm, one that people follow only "when they perceive it to be in their interest to do so," will be followed less frequently in the population than a norm that has been internalized (Gintis, 2003 408). He argues that cooperation enhancing norms, like altruistic punishment, are internalized through our culture and our genetics.

While the Gintis (2000) model showed that altruistic punishment is an evolutionarily stable strategy, this model was "sensitive to group size and migration rates" (Gintis, 2003 416). However, "the gene-culture coevolutionary model presented in this paper" is not so sensitive because "the fitness costs of altruistic punishment are low" and therefore "a replicator dynamic is unlikely to render the altruism equilibrium unstable in this case" (Gintis, 2003 416). Accordingly, he has developed a more sophisticated model of strong reciprocity, independent from "repeated interaction, reputation effects, or multi-level selection" (Gintis, 2003 416). While

Gintis acknowledges that his model has faults, for example that payoffs remain fixed when in fact "the payoff to being self-interested may increase when agents are predominately altruistic," nonetheless it is less sensitive to replicator dynamics and material incentives, like reputation effects (Gintis, 2003 416)

His model presents a powerful new way of looking at the origins of strong reciprocity. He shows that human culture allows us to internalize norms with "the capacity to enhance the fitness of the individuals who express them" (Gintis, 2003 417). The evolution of altruistic punishment may seem paradoxical because it is individually fitness reducing. Yet human cultural evolution occurred in tandem with human genetic evolution. Gintis' models shows that "altruistic norms can hitchhike on personally fitness-enhancing norms" (Gintis, 2003 417). Accordingly, humans develop a normative preference for altruism because these genetic "hitchhiker norms" that enhance group fitness are indistinguishable from those that merely enhance our individual fitness. If altruistic norms did not attach to fitness-enhancing norms, "human society as we know it would not exist" (Gintis, 2003 417). The evolution of altruistic punishment is logical: "it is generally prudent to develop a reputation for punishing those who hurt us" (Gintis, 2003 418). In terms of culture, it would be beneficial if all individuals were strong reciprocators. Thus, "it is a short step to turning this prudence into a moral principle," a normative claim that will be socially internalized by all (Gintis, 2003 418).

Criticisms of Group Selection Models

The proponents of group selection theories acknowledge shortcomings in their own models. One shortcoming that is common to all public goods games is that the payoff disadvantage of punishers relative to contributors may not decrease. Altruistic punishment can evolve from a group-selection model if the punishers' payoff disadvantage relative to

contributors becomes negligible as fewer individuals imitate the defection strategy. However, the punishers may still have a fitness disadvantage if mutations can occur. For example, "the costs of monitoring or punishing occasional mistaken defections would mean that punishers have slightly lower fitness than contributors" (Boyd et al. 3531). When punishers' relative fitness disadvantage is maintained, then no one will imitate altruistic punishers and accordingly no one will have an incentive to cooperate. With payoff asymmetries, "defection is the only one of these three strategies that is an evolutionarily stable strategy in a single isolated population" (Boyd et al. 3531). Therefore, if monitoring costs are high or "when the probability of mistaken defection is high enough that punishers bear significant costs even when defectors are rare, group selection does not lead to the evolution of altruistic punishment" (Boyd et al. 3533). The success of group-selection models is contingent on relatively low costs for monitoring and even lower likelihood of mutation or mistakes. While perhaps it is unrealistic to expect such limitations, the models nonetheless sustain how altruistic punishment could have evolved under these conditions in early human history.

Next, theories of group selection rely on the existence of substantive differences between groups of humans. Cooperation evolves because altruistic punishment "in combination with the imitation of economically successful behaviours prevents the erosion of group differences with regard to the relative frequency of cooperation members" (Fehr and Fischbacher 790). Cooperative groups have higher payoffs, survive and are imitated whereas non-cooperative groups, without punishers or high frequencies of defectors, become extinct. These differences between groups determine their relative fitness. However, if groups are alike in terms of a certain feature, then that feature cannot be the mechanism that gives one group a fitness advantage over the others.

Critics provide several reasons for why the substantive differences between human groups that are needed to sustain group selection do not exist. The first criticism is that humans are extremely genetically similar. Therefore, "Multilevel selection theories only provide plausible ultimate explanations of human altruism, however, if they are interpreted in terms of cultural evolution rather than genetic evolution…because cultural variation between groups is much bigger than the genetic variation between groups" (Fehr, Fischbacher and Gächter 5). This criticism allows the argument that cultural variation may be substantive enough to sustain group selection. Where group norms of cooperation and punishment differ between groups, group selection theories could still be valid. Moreover, these critics argue that cultural variation is relatively bigger than genetic variation. There are two problems with this criticism. While cultural variation may be larger than genetic variation, this does not mean that genetic variation could not be responsible for group selection; instead, it simply means that evolutionary selection forces acting on cultural variation could be stronger. Yet it does not exclude the possibility of selection acting on genetic differences, albeit small ones. Second, the criticism assumes that the genetic variation among human groups is small. In terms of modern day humans, this is undoubtedly true. Yet one can easily imagine early stages of human development in which isolated communities had vastly different genetic features, such as alleles, from neighboring groups. Because altruistic punishment has evolved throughout the course of human history, the contention that there could not have been large genetic variations among isolated groups is dubious.

The second criticism is that humans are not confined to social groups: they can leave one group and join another. This migration "between groups removes the differences between groups" (Fehr and Fischbacher 64). Fehr and Fischbacher provide an example of a potentially

destabilizing invasion. Migrant defectors can invade a society of altruistic groups. The defectors have a fitness advantage and "will reproduce at a higher rate, quickly removing the differences in the composition of selfish and altruistic individuals across groups. Thus, group selection cannot become operative" (Fehr and Fischbacher 64). However, defectors will only have a higher relative fitness if they invade altruistic communities that do not punish. Accordingly, "selfish migrants may not be able to reproduce at a higher rate in the presence of social norms proscribing individually selfish behavior because they are punished for violation of the norm" (Fehr and Fischbacher 64). Where altruistic punishment has sustained a cooperative equilibrium, defectors do not have a relative fitness advantage and variation between groups remains constant. Thus, group selection theory is unaffected by this criticism.

A similar argument is made about the invasion of free-riders. Because contributors who do not punish have a fitness advantage relative to punishers, they can invade a cooperative equilibrium and outcompete punishers. Once punishers have gone extinct, defectors can invade and the cooperative equilibrium will be obsolete. However, Henrich and Boyd argue that this infinitely regressive effect will never occur. While there is a payoff asymmetry between altruistic punishers and cooperators, this asymmetry only exists when there are defectors. Even if cooperators come into a population with punishers, defection will not increase. The existence of punishers means that "defection does not pay" and "the only defections will be due to rare mistakes, and thus the *difference* between the payoffs of punishers and second-order free-riders will be relatively small" (Henrich and Boyd 81). This "anti-social invasion" of non-punishing contributors "may eventually destabilize cooperation" if there is a huge probability of mutation such that all punishers must punish defectors, creating an absolute payoff disadvantage for punishers relative to cooperators (Henrich and Boyd 88). This will be determined by the

dynamics of the model. When such a high rate of mutation is unlikely, migrant cooperators will not outcompete punishers and defectors will never be given the opportunity to invade a cooperative group so long as altruistic punishment is the norm. This maintains a substantial difference in the fitness between cooperative and non-cooperative groups and thus group selection can sustain the cooperative equilibrium even if there are migrations between groups.

Another argument is that theories of human cooperation based on punishment are so radically different from other species' behavior that the natural world does not give credence to these purportedly biological explanations. Critics contend that the mechanism and behaviors suggested for sustaining cooperation, i.e. strong reciprocity, "are seldom observed in other animals" (Bowles and Gintis 25). Bowles and Gintis respond that this is because their model, and others like it, rely on "cognitive, linguistic, and other capacities unique to our species" (Bowles, Gintis 25). For example, Gintis (2000) argues that "as a result of the superior tool-making and hunting ability of *Homo Sapiens*, the ability to inflict costly punishment (high h) at a low cost to the punisher (low c_r), probably distinguishes humans from other species that live in groups," thus allowing human strong reciprocators to have low costs of punishment (Gintis, 2000 174). Similary, Bowles and Gintis defend ostracism as a punishment strategy by suggesting that "uniquely human capacities to inflict punishment at a distance, through projectile weapons, reduce the cost of ostracizing a norm violator" (Bowles and Gintis 26). Moreover, they contend that "strong reciprocity emerged through a modification of reciprocal altruist behaviors;" because "reciprocal altruism appears to be very rare in other species," they postulate that strong reciprocity might also be an exclusively human evolutionary adaptation (Bowles and Gintis 26).

While many of the criticisms against group selection theory can be answered, other arguments in recent literature make a stronger case. Gardner and West (2004) argue that

scientists have frequently rejected "kin selection" arguments—they recognize that "relatedness is too low" for this strategy to be stable in large groups—yet "group selection has often been regarded as important" (Gardner and West 761). They continue, "Kin selection and group selection are mathematically equivalent ways of conceptualizing the same evolutionary process" (Gardner and West 761). Thus, the dichotomy that previous researchers have constructed between kin and group selection is a false one. Further, the kin/group selection strategy is not sufficient to explain the evolution of altruistic punishment because the theory depends on so many contingencies: kin/group selection only functions "insofar as the benefit to the group is large enough, the cost to the individual is low enough, and there is substantial between-group as opposed to within-group variation in trait values" (Gardner and West 761). Gardner and West "link kin selection, group selection, and cultural group selection in terms of a generalized view of relatedness" and propose an alternative to this corpus of research (Gardner and West 754).

Altruistic punishment could evolve as a stable strategy "in the absence of relatedness, partner recognition, reputation, and any mechanism whereby an individual may bias her interactions or tailor her behavior in response to her immediate social partner" (Gardner and West 762). It is not the genetic, cultural or individual relationship between individuals "that facilitates the evolution of punishing behavior. What is crucial is that there is a positive correlation between the punishment strategy played and cooperation received by an individual" (Gardner and West 754). Punishing behavior would not have developed or become evolutionarily stable without this positive association (Gardner and West 755-7).

The punishment strategy is costly because "punishment acts to directly reduce both the fitness of the actor and the fitness of her social group" (Gardner and West 755). While cooperation maintains or increases group fitness, punishment actually reduces group fitness. Yet

punishment is considered altruistic because it "indirectly" benefits the group by creating "a coercive social environment in which cooperation is favored" and thus "protects the social group from the breakdown of cooperation" (Gardner and West 761). Individuals will only punish when they receive the benefit of cooperation in return; if "individuals facultatively adjust their level of cooperation in response to the local threat of punishment," or cooperate when threatened with punishment, then "full punishment can be an evolutionarily stable strategy" (Gardner and West 760).

Gardner and West do not deny that some features of kin/group selection, including relatedness and size, would facilitate the evolution of cooperation and punishment. If all of the individuals in a group are punishers, for example "in a viscous population where genealogical kin tend to associate with each other," then the group members are assured of a positive association between punishing others and receiving cooperation in return (Gardner and West 762). Therefore, cooperation could have originated in "altruism between relatives" followed by the evolution of punishment "to favor and maintain higher levels of cooperation" amongst unrelated individuals (Gardner and West 761). Likewise, punishment may have evolved within "small groups of interacting individuals," a "social structure" that is "more conducive" to punishment; "once common, punishment could be retained even when interaction began to occur within much larger groups of humans" (Gardner and West 761).

However, Gardner and West caution that the mechanism behind the continuation of punishment beyond small, related groups is not the kin selection mechanism that allowed cooperation and punishment to evolve within these groups. Instead, they believe that punishment expands because of "niche construction," meaning that the punishing behavior "modifies the social environment in such a way as to alter the selective pressures acting upon other traits"

(Gardner and West 762). At the individual level in much larger groups, the individual has no incentive to deviate form a punishing strategy if "punishment is already frequent" because "the fitness saved by forgiving is minimal and may be overwhelmed by the concomitant decline in the amount of cooperation received because of the decrease in selection for cooperation among social partners" (Gardner and West 762). Thus, social partners will reject an individual who does not punish if punishment has already evolved as the social norm. Accordingly, an individual will cooperate and punish when both are established norms in order to gain "the direct benefits accrued when cooperation is facultative" (Gardner and West 762). In tandem, the benefits of cooperating and costs of rejection by potential social partners are sufficient selection pressures "to maintain punishment among humans, rendering elaborate population dynamics and cultural practices unnecessary" (Gardner and West 762)

While the individual level selection theory reduces the contingencies required for the evolution of altruistic punishment in a kin/group selection model, the Gardner and West model has problems of its own. They emphasize the importance of the positive association between punishing others and receiving cooperation in return yet determining this association "could be hard to test directly, especially experimentally, because of limitations on how an individual's level of punishment could be manipulated" (Gardner and West 761). Their model works in theory but may be more difficult to prove in laboratory experiments. Second, they emphasize that individual level selection makes group level selection obsolete; yet "numerical analysis of the example model reveals that increasing the frequency of maladaptive behavior reduces the likelihood that individual level selection will be able to maintain altruistic punishment in very large groups" (Gardner and West 762). This means that as more individuals behave asocially—i.e. by cooperating and punishing in response to the negative selection pressures associated with

defecting—altruistic punishment could not have evolved through the individual selection mechanism.

At worst, even if researchers reject the individual selection theory because of the difficulties associated with asocial behavior, Gardner and West have at least challenged the dominant acceptance of the kin/group selection theory as the explanation for the evolution of altruistic punishment. In their model, the individual fitness benefit of instigating cooperation by punishing others is the mechanism that allowed punishment to evolve and become stable. Their theory emphasizes the importance of the human ability to adapt behavior for particular social environments, an emphasis on individual agency lost in the group mentality of other theories.

I now turn to a discussion of asocial behavior. This criticism of the evolution of altruistic punishment and strong reciprocity is that scientists assume that all individuals will want to participate in the public goods game. Hauert, De Monte, Hofbauer, and Sigmund (2002), Fowler (2005) and Brandt, Hauert and Sigmund (2006) all present theories in which individuals have the option not to participate.

Hauert, De Monte, Hofbauer, and Sigmund agree that humans may want to participate in public goods games in order to receive benefits, and can be motivated to cooperate through the rewards and punishments of others. Yet they propose that cooperation can be achieved without rewards and punishment. An individual could be a loner; instead of participating in the public goods game, she chooses "to fall back on a safe 'side income' that does not depend on others. Such risk-averse optional participation can foil exploiters and relax the social dilemma, even if players have no way of discriminating against defectors" (Hauert, De Monte et al. 1129).

Cooperating in the public goods game will redound to an individual's benefit unless "defectors are prevalent;" many defectors reduce the payoffs from participating in the public

goods game such that "it is better to stay out of the public goods game and resort to the loners' strategy" (Hauert, De Monte et al. 1130). However, when loners are prevalent, cooperators become successful by forming "groups of small size *S*" (Hauert, De Monte et al. 1130). Cooperation pays, despite the existence of defectors and loners: "Although defectors always do better than cooperators, in any given group, the payoff for cooperators, when averaged over all groups, will be higher than that of defectors (and loners), so cooperation will increase" (Hauert, De Monte et al. 1130). While defection is the dominant strategy in large groups, cooperation is dominant in small groups, and "mere option to drop out of the game preserves the balance between the two options, in a very natural way" (Hauert, De Monte et al. 1130). The dynamics of the game changes depending on what strategy each player adopts, e.g. imitate-the-best (Hauert, De Monte et al. 1130). Even if the dynamics indicate that defection reaches fixation in the public goods game, "the drop-out option allows groups to form on a voluntary basis and thus to relaunch cooperation again and again" (Hauert, De Monte et al. 1131). As groups grow larger, however, the individuals are incentivized to drop because of "an increased threat of exploitation;" thus, "individuals keep adjusting their strategies but in the long run do no better than if the public goods option had never existed" (Hauert, De Monte et al. 1131). However, the drop out option, or voluntary participation, "avoids the deadlock of mutual defection that threatens any public enterprise in larger groups" (Hauert, De Monte et al. 1131). Thus, Hauert, De Monte et al. propose a solution to the public goods game's free-rider dilemma but present a new puzzle: individuals are no better off when they cooperate in the public goods game if a loner strategy offers greater payoffs.

Fowler's model is similar to the preceding model, yet allows for the existence of punishers alongside cooperators, defectors and loners. Punishers dole out punishment by

assessing the standing of other members: cooperators achieve "good standing," defectors achieve "bad standing" and loners "avoid a bad standing designation by not participating. This feature of the model prevents defectors from completely taking over the population because they are susceptible to nonparticipants" (Fowler 7048). His model's payoffs are such that punishing second-order free-riders, those who contribute in the public goods game but do not punish, "can be small or infrequent" because any punishment greater than zero "gives punishers an advantage over contributors" (Fowler 7048). In addition, his model stipulates that a group of punishers whose punishment is less costly cannot invade a population of punishers because "punishers also punish anyone who does not punish nonpunishers enough" (Fowler 7048). Fowler shows that defectors, who reduce the amount of the public good and accordingly their own income, will do worse than "nonparticipants who rely on their own activities" (Fowler 7048). Consequently, "cooperation-enhancing strategies like altruistic punishment have an opportunity to evolve because they simultaneously acquire more benefits than nonparticipants and keep defectors at bay" (Fowler 7048).

This model shows that altruistic punishment can evolve in a population in which both contribution and punishment are dominated strategies and that "the origin and persistence of widespread cooperation is possible with voluntary, decentralized, anonymous enforcement, even in very large populations under a broad range of conditions" (Fowler 7048). Like Hauert, De Monte, Hofbauer and Sigmund, Fowler's model shows that there is a "cycle of cooperation, defection, and nonparticipation," depending on population dynamics (Fowler 7048). He thinks that this cycle "is important for understanding the origin of cooperation but may not be useful for understanding its persistence. When altruistic punishment evolves, the cycle should disappear and cease to be observed in the population dynamics" (Fowler 7048). Otherwise, altruistic

punishes causes the cycle to reach fixation rather than solely cooperation. Lastly, his "model suggests that there are restrictions on what kinds of strategies punishment can evolve" and that punishment will not evolve strategies "that yield a payoff disadvantage" to any individual (Fowler 7048).

In response to Fowler's argument, Brandt, Hauert, and Sigmund (2006) show that punishment-induced cooperation is not the only Nash equilibrium for a public goods game; instead, both punishing in and abstaining from a public goods game "are possible as long-term outcomes" (Brandt et al. 497). Like the other models, their model gives individuals the option to "opt out" of the public goods game altogether. These loners exist apart from the cooperators, defectors and punishers in the public goods game, obtaining "an autarkic income independent of the other players' decision" (Brandt et al. 495). To reiterate, a loner's income is independent of the public goods game, so she is not a free-rider. They conclude that in contrast to Fowler, "our model displays a bistable behavior" (Brandt et al. 496). The dynamics depends on the initial concentration of each type of player, yet the model always converges to one of two equilibria: "either to a Nash equilibrium consisting of cooperators and punishers, or to a periodic orbit in the face w= *0* (no punishers), where the frequencies of loners, defectors, and cooperators oscillate endlessly" (Brandt et al. 496). Furthermore, Fowler's model is biased toward the evolution of altruistic punishment as the only equilibrium. First, Fowler's model allows punishers to punish cooperators "even if there are no defectors around, and thus [cooperators] will be unable to invade a population of punishers by neutral drift" (Brandt et al. 496-7). Fowler's attempt to solve the second order free-rider problem means that punishers will punish cooperators for not punishing even when there are no defectors around to punish. Fowler's model therefore unnecessarily reduces the fitness of cooperators, which precludes the possibility of a second

equilibrium. Second, Fowler does not consider the absence of punishers. Brandt, Hauert and Sigmund argue that without punishment, "each invasion of contributors is quickly repressed so that, up to rare, intermittent bursts of cooperation, the population is reduced to the autarkic way of life;" thus, the second equilibrium of all loners, or all abstain from the public goods game, is never considered in Fowler's model (Brandt et al. 496-7).

In sum, the results of these three models present an interesting new interpretation of the potential for the evolution of cooperation in public goods games. Hauert, De Monte, Hofbauer, and Sigmund (2002) and Fowler (2005) both present the concern that when participants are given the option to opt out of the public goods game, the dynamics can create a cycle of defection and nonparticipation. While altruistic punishment can stabilize cooperation if there are punishers, cooperation is only sustained for a fraction of the cycle. This means that altruistic punishment will at best conditionally stabilize cooperation. Brandt, Hauert and Sigmund (2006) present the worst case scenario, an equilibrium in which the autarkic way of life dominates cooperation in the absence of punishers. Fowler and Brandt, Hauert and Sigmund's theories show that the success of altruistic punishment in upholding cooperation is contingent on the proportion of punishers that exist in the initial game. When people have the option to leave the public goods game, altruistic punishment will only be stable if enough people are willing to incur a personal cost to procure a social good. We might be able to assume that for our earliest ancestors this premise was true, given modern man's tendency to punish altruistically in public goods games.

Discussion

The group selection models with gene-culture co-evolution provide justification for the thesis that the evolution of strong reciprocity can stabilize cooperative equilibriums in public goods games. While the researchers acknowledge that their models depend on dynamics and

contingencies, like population size or mutation rate, it is reasonable to believe that these dynamics could have existed at early stages of human development. If we accept the Gardner-West premise that group-selection models are glorified kin-selection theory, their model explains how altruistic punishment could become an evolutionarily stable strategy through a mechanism other than group-selection. Moreover, laboratory experiments confirm that humans behave like strong reciprocators and have neurological incentives to punish others. Our evolutionary past must provide some justification for our modern day behavior and motivations.

However, there are legitimate doubts as to whether the evolutionary explanation for altruistic punishment is valid. The asocial evolutionary theories represent a true challenge to the evolution of altruistic punishment in that they show that the presence of altruistic punishers may not necessarily sustain cooperation. The pro-social behaviors we see today might be explained by an alternative theory of behavior that consistently secures cooperation. Moreover, some researchers contend that the behavior observed in public games experiments in laboratory settings do not represent actual human behaviors. Strong reciprocity in one-shot, anonymous interactions is an experimental fiction. Arguments for the evolution of strong reciprocity are nullified if the behavior they seek to explain does not actually exist.

Some researchers argue that our evolutionary past *precludes* the possibility of distinguishing a laboratory setting from a social dilemma. For example, Cultural anthropologists and evolutionary psychologists claim that "in the environment of evolutionary adaptation (EEA) or ancestral past, people mostly engaged in repeated games with people they knew. Evolution created specialized cognitive heuristics for playing repeated games efficiently" (Fehr, Fischbacher and Gächter 18). Neurologically, a test subject's brain "is not a general purpose information processor, but rather a set of interacting modular systems adapted to solving the

particular problems faced by our species in its evolutionary history" (Gintis et al. 168). The theory that the laboratory is an "unnatural habitat," an alien landscape that prevents subjects from responding appropriately, "assumes the *absence* of a module or cognitive heuristic which could have evolved but did not—the capacity to distinguish temporary one-shot play from repeated play" (Fehr, Fischbacher and Gächter 19). They continue that "the anonymous nonrepeated interactions characteristic of experimental games were not a significant part of our evolutionary history;" accordingly, humans would not have adapted to this type of situation and cannot "behave in a fitness-maximizing manner" without an evolutionary adaptation to the laboratory (Gintis et al. 168). Thus, the subjects behave in the laboratory as if the experiment was "a nonanonymous, repeated interaction" and "maximize fitness with respect to this reinterpreted environment" (Gintis et al. 168). The results from public goods experiments cannot claim that individuals will behave altruistically in one-shot, anonymous interactions because humans *never* act as if a social situation is constrained in this way.

Proponents of strong reciprocity think the unnatural habit theory is fallacious. One response is that "even if strong reciprocity were a maladaptation, it could nevertheless be an important factor in explaining human cooperation today" because modern society requires us to cooperate with other people who we may never see again (Gintis et al. 168). Also, they do not agree that test subjects will confuse the laboratory setting with non-anonymous, repeated interaction. They believe that "humans are well capable of distinguishing individuals with whom they are likely to have many future interactions" and will "cooperate much more if they expect frequent future interactions than if future interactions are rare" (Gintis et al. 168-9). Public goods data from Fehr and Gächter (2000) indicates that while subjects cooperate in various permutations of the public goods games, "cooperation rates are generally lower in public good

games when the group composition changes randomly in very period than they are when the group composition is constant across all ten periods. This fact suggests that, on average subjects can distinguish between one-shot and repeated interactions" (Fehr, Fischbacher and Gächter 19). Thus, empirical data denies the unnatural habitat conclusion. The human brain is not so intimately tied to evolutionary conditions that test subjects cannot understand the laboratory context.

I believe that in light of the unnatural habitat critique, more attention needs to be paid to experimental controls and constraints. First, "a fully satisfactory test of subjects' capacity to distinguish one shot from repeated interactions requires that the same subjects participate in both conditions so that we can examine behavioral changes across conditions at the individual level" (Fehr, Fischbacher and Gächter 19). This means that subjects must participate in the Partner and the Stranger treatments as well as in one-shot and iterated games before further conclusions can be drawn. Controlling for all of these factors, and evaluating test subjects' survey responses, will confirm the existence of strongly reciprocal behavior. Second, future experimenters need to consider the likelihood that test subjects may interact in the future. Fehr and Schmidt (1999) argue that "the social context and the institutional environment in which interactions take place is likely to be important" and have an influence on experimental results (Fehr and Schmidt 851). For example, many experiments use students that attend the same University. Ethnographical studies, especially in developing societies, focus on one particular ethnic group. The likelihood that these players will interact in the future is high precisely because their campus or social community has a finite spatial demarcation. Thus, these subliminal considerations may shift the altruistic tendencies of the players toward a more cooperative outcome.

Another argument is that cooperation is only sustained in public goods games because of symmetric payoffs. If people have evolved to internalize equitable norms, then they will care very much if the outcome of a game is fair. However, they may not participate in a public goods game if it favors certain individuals such that the potential for an equitable outcome is obsolete. Fehr and Schmidt argue: "it will be more difficult to sustain cooperation if the game is asymmetric. For example, if the public good is more valuable to some of the players, there will in general be a conflict between efficiency and equality" (Fehr and Schmidt 846). They contend that even if players can punish others, when "the game is sufficiently asymmetric it is impossible to sustain cooperation" (Fehr and Schmidt 846). Experimental and evolutionary models that assume that all agents contribute to a public good in order to receive the same payoffs in monetary units or improved fitness may need to revise their results if the payoff opportunities are not the same for everyone. I believe that if a public goods game has an institutionalized discrimination against rewarding all players equitably, then those players who would not be rewarded may opt out of the public goods game. Altruistic punishment may fail to provide a cooperative equilibrium if a sufficient number of individuals sense that there will be a payoff asymmetry. This could be an interesting theory to pursue in light of modern discourses on institutional discrimination against social "others," such as women or ethnic minority groups. I believe that future researchers should carefully consider the institutional payoff asymmetry critique.

Conclusion

Current research has set an important precedent in uncovering the evolutionary origins of cooperation in public goods games, particularly when cooperation is sustained by altruistic punishment. I think that the experimental data strongly suggests that humans have an inclination

toward altruistic punishment; the neurological results all but confirm that humans are rewarded when they sustain cooperative equilibria by punishing others. There must be an evolutionary explanation for this neurological response. Arguments for the evolutionary internalization of cultural norms of collaboration, cooperation and equity are on the right track; these values are dominant in many modern human societies. Moreover, the notion that pro-sociality may have become ingrained in the human genetic makeup answers the question of psychological motivations for altruistic punishment.

However, group selection models may be unnecessarily complex and even inaccurate in explaining how altruistic punishment can sustain cooperation. Further research must be done, not only in better controlled laboratory experiments but also with models that are independent of dynamics and consider opting out of the public goods game. There is certainly evidence that exploration of strong reciprocity is a promising venture for game and evolutionary theorists alike. Future analysis will hopefully confirm that human cooperation can find its origins in the selfless punishment of others.

End Notes

[i] Public goods games have been described in the literature using a variety of names: "Tragedy of the Commons, Free Rider Problem, Social Dilemma, or Multiperson Prisoner's Dilemma—the diversity of the names underlines the ubiquity of the issue" (Hauert, De Monte et al. 1129). I will call this cooperative dilemma a public goods game throughout the paper for uniformity.

[ii] Unless otherwise indicated, text that is italicized in citations was italicized in the original document.

[iii] Some researchers do not use the terms interchangeably. They believe that strong reciprocators have selfish motives that "induce them to increase rewards and punishment in repeated interactions or when reputation-building is possible" (Fehr and Fischbacher 788). However, in the context of public goods games, "rewarding" behavior is simply cooperating by contributing the expected amount to the public good. An altruistic punisher will contribute to the public good, or engage in this same "rewarding behavior," so long as she does not play the hypocritical strategy, which I am not considering. Moreover, I am excluding material incentives, such as reputation building, for punishing strategies precisely because I do not want to consider selfish motivations. I am therefore saying that altruistic punishment and strong reciprocity are synonymous to the extent that strongly reciprocal behavior is altruistic and punishers contribute to the public goods game.

References

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

Bowles, Samuel and Herbert Gintis, 2004. "The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations." *Theoretical Population Biology*, 65: 17-28

Boyd, Robert, Herbert Gintis, Samuel Bowles, and P.J. Richerson, 2003. "The Evolution of altruistic punishment." *Proceedings of the National Academy of Sciences (USA)*, 100: 3531-3535.

Brandt, Hannelore, Christoph Hauert and Karl Sigmund, 2006. "Punishing and Abstaining for public goods." *Proceedings of the National Academy of Sciences (USA*), 103:495-7.

De Quervain, Dominique J.-F., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrico Schnyder, Alfred Buck and Ernst Fehr, 2004. "The Neural Basis of Altruistic Punishment." *Science*, 305(5688): 1254-8.

Fehr, Ernst, Urs Fischbacher and Simon Gächter, 2002. "Strong Reciprocity, Human Coopreation, and the Enforcemtn of Social Norms." *Human Nature*, 13(1): 1-25.

Fehr, Ernst and Urs Fischbacher, 2003. "The Nature of Human Altruism." *Nature*, 425: 785-91.

Fehr, Ernst and Urs Fischbacher, 2004. "Third-party punishment and social norms." *Evolution and Human behavior*, 25: 63-87.

Fehr, Ernst and Simon Gächter, 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, American Economic Association, 90(4): 980-994

Fehr, Ernst and Simon Gächter, 2002. "Altruistic Punishment in Humans." *Nature*, 415: 137-140.

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

Fowler, James H., 2005. "Altruistic Punishment and the Origin of Cooperation." *Proceedings of the National Academy of Sciences (USA)*, 102(19): 7047-49

Gardner, Andy and Stuart A. West, 2004. "Cooperation and Punishment, Especially in Humans." *The American Naturalist*, 164 (6): 753-764.

Gintis, Herbert, 2000. "Strong Reciprocity and Human Sociality." *Journal of Theoretical Biology*, 206: 169-79.

Gintis, Herbert, 2003. “The Hitchhiker’s Guide to Altruism: Gene-culture Coevolution and the Internalization of Norms.” *Journal of Theoretical Biology*, 220: 407-418.

Gintis, Herbert, Eric Alden Smith and Samuel Bowles, 2001. “Costly Signaling and Cooperation.” *Journal of Theoretical Biology*, 213: 103-119.

Hauert, Cristoph, Silvia De Monte, Josef Hofbauer and Karl Sigmund, 2002. “Volunteering as Red Queen Mechanism for Cooperation in Public Goods Games.” *Science*, 296: 1129-1132.

Heckathorn, Douglas D, 1989. “Collective Action and the Second-Order Free-Rider Problem.” *Rationality and Society,* 1(1): 78-100.