# Behavioral Ethics

Herbert Gintis[*]

September 6, 2009

## 1 Behavioral Ethics and the Rational Actor Model

Philosophical views about morality have traditionally been supported by abstract reasoning and introspection, with at best passing reference to actual human behavior. Behavioral ethics develops models of human morality based upon the fact that morality is an emergent property of the evolutionary dynamic that gave rise to our species. Propositions concerning moral behavior are framed and tested using the methods of game theory, using subjects from a variety of social backgrounds and cultures. In this paper I will outline some major themes in behavioral ethics, and suggest how they relate to philosophical ethics.

*Homo sapiens* is one of many social species, but alone is the product of an evolutionary dynamic known as *gene-culture coevolution*. Like language and the social emotions (shame, pride, contempt, empathy) morality is an emergent human universal.

Individual fitness in humans depends on the structure of social life. Because culture is limited by and facilitated by human genetic propensities, human cognitive, affective, and moral capacities are the product of an evolutionary dynamic involving the interaction of genes and culture (Cavalli-Sforza and Feldman 1982, Boyd and Richerson 1985, Dunbar 1993, Richerson and Boyd 2004). This coevolutionary process has endowed us with preferences that go beyond the self-regarding concerns emphasized in traditional economic and biological theory. Gene-culture coevolution explains why we have a social epistemology facilitating the sharing of intentionality across minds, as well as why we have such non-self-regarding values as a taste for cooperation, fairness, and retribution, the capacity to empathize, and the ability to value character virtues (e.g., honesty, loyalty, trustworthiness).

[*] Santa Fe Institute and Central European University. To appear in Edward Slingerland and Mark Collard (eds), *Integrating Science and the Humanities: Interdisciplinary Approaches*. I would like to thank Edward Slingerland for helpful comments on a draft of this paper.

In behavioral ethics, we recognize that people consider moral statements to have truth values, but we consider these values as being valid only for the specific social group involved, rather than having universal scope. We thus treat ethics in a manner similar to linguistics, where grammaticality and correct usage are important and analytically tractable, yet highly specific to a particular society of speakers. As it turns out, there are a host of human values that are universal in that they are present in virtually every known society (Brown 1991). We account for this universality in terms of the general requirements for social living of our species.

Moral behavior is often held to be incompatible with rational choice. This is incorrect. The rational actor model of economic theory presupposes that people have consistent preferences, but does not require that preferences be self-regarding or materialistic. We can just as easily chart how people value honesty or loyalty in the same way that we can chart how they value fried chicken or cashmere sweaters.

Some caveats are in order, however, in interpreting the rational actor model. First, individuals do not consciously maximize "utility," or anything else. As long as people have consistent preferences, we can represent their choices as the solution to a maximization problem. This is analytically useful and gives accurate predictions, but it does not mean that individuals really "maximize." This analytical ploy is similar to the physicist predicting the path followed by a light wave by assuming the light minimizes path length traveled. This works, but of course light waves do not really minimize anything.

Second, individual choices, even if they are self-regarding (e.g., personal consumption) are not necessarily welfare-enhancing. In the sense of the rational actor model, it may thus be rational to smoke, have unsafe sex, and even cross the street without looking. Third, individual preferences are not fixed, but rather are a function of an individual's current state (e.g., state of hunger or sexual arousal). Fourth, the fact that we model individuals as rational decision-makers does not imply that we are methodological individualists. In particular, we may adopt the rational actor model yet assert that is not possible to derive either social norms or moral values from individual rationality. In fact, it is quite possible that both social norms and the moral realm are emergent properties of human evolution that cannot be derived from the aggregation of individual rational choice.

Beliefs are the Achilles' heel of the rational actor model. The standard model of rational choice treats beliefs as purely subjective (the so-called *subjective priors*), whereas individual beliefs are in fact a part of a social network of interdependent socially constructed beliefs. The rational actor model is really a lower-dimensional projection of a general model of human action, which includes an interpretive and a deliberative dimension as well. Full human reason operates in this larger action framework.

Because the use of the word "rational" in the rational actor model is so circumscribed compared with the general usage of the word, we often call the rational actor model the *beliefs, preferences, and constraints* model (BPC), because this capture the notion of consistent preference, the centrality of beliefs, and the notion making tradeoffs subject to informational and material constraints.

In the BPC model, choices give rise to probability distributions over outcomes, the expected values of which are the payoffs to the choice from which they arose. Game theory extends this analysis to cases where there are multiple decision makers. In the language of game theory, players are endowed with strategies, and have certain information, and for each array of choices by the players, the game specifies a distribution of payoffs to the players. Game theory predicts the behavior of the players by assuming each is rational; in other words, each maximizes a preference function subject to beliefs as well as informational and material constraints.

The experiments described below are all based on using game theory to set up the choices available to subjects, the knowledge they have on which their choices are based, and the payoffs to each subject as a function of their joint strategy choices. We assume the subjects are rational (i.e., consistent) decision-makers, so that their choices reflect their subjective trade-offs among heterogeneous payoffs—some material and some moral and/or other-regarding.

## 2 Experimental Findings on the Rationality of Altruistic Behavior

There is nothing irrational about caring for others. But do preferences for altruistic acts entail transitive preferences as required by the notion of rationality in decision theory? Andreoni and Miller (2002) showed that in the case of the Dictator Game, they do. Moreover, there are no known counterexamples.

In the Dictator Game, the experimenter gives a subject, called the Dictator, a certain amount of money and instructs him to give any portion of it he desires to a second, anonymous, subject, called the Receiver. The Dictator keeps whatever he does not choose to give to the Receiver. Obviously, a self-regarding Dictator will give nothing to the Receiver. Suppose the experimenter gives the Dictator $m$ points (exchangeable at the end of the session for real money) and tells him that the price of giving some of these points to the Receiver is $p$, meaning that each point the Receiver gets costs the giver $p$ points. For instance, if $p = 4$, then it costs the Dictator 4 points for each point that he transfers to the Receiver. The Dictator's choices must then satisfy the budget constraint $\pi_s + p\pi_o = m$, where $\pi_s$ is the amount the Dictator keeps and $\pi_o$ is the amount the Receiver gets. The question, then, is simply, is there a preference function $u(\pi_s, \pi_o)$ that the Dictator maximizes subject to the budget constraint $\pi_s + p\pi_o = m$? If so, then it is just as rational, from a behavioral standpoint, to care about giving to the Receiver as to care about consuming marketed commodities.

Varian (1982) developed a generalized axiom of revealed preference (GARP) that ensures that individuals are rational as in the sense of traditional consumer demand theory. Andreoni and Miller (2002) worked with 176 students in an elementary economics class and had them play the Dictator Game multiple times each, with the price $p$ taking on the values $p = 0.25, 0.33, 0.5, 1, 2, 3$, and 4, with amounts of tokens equaling $m = 40, 60, 75, 80$, and 100. They found that only 18 of the 176 subjects violated GARP at least once and that of these violations, only four were at all significant. By contrast, if choices were randomly generated, we would expect that between 78% and 95% of subjects would have violated GARP.

As to the degree of altruistic giving in this experiment, Andreoni and Miller found that 22.7% of subjects were perfectly selfish, 14.2% were perfectly egalitarian at all prices, and 6.2% always allocated all the money so as to maximize the total amount won (i.e., when $p > 1$, they kept all the money, and when $p < 1$, they gave all the money to the Receiver).

We conclude from this study that, at least in some cases, and perhaps in all, we can treat altruistic preferences in a manner perfectly parallel to the way we treat money and private goods in individual preference functions. We use this approach in the rest of the problems in this chapter.

### 2.1 Conditional Altruistic Cooperation

A *social dilemma* is a situation in which a number of people can gain by cooperating, but cooperating is costly, so each individual does better personally by not cooperating, no matter what the others do. For instance, suppose if a member of a group of size $n \geq 2$ pays the cost $c > 0$, he benefits each of the others by an amount $b > 0$. If $b(n-1) > c$, we have a social dilemma: at cost $c$, an individual can help the group by the amount $(n-1)b > c$, but a selfish individual will not do so. If all cooperate, each will earn $b - c > 0$, but in a group of self-regarding individuals, each will earn zero.

*Conditional altruistic cooperation* is a predisposition to cooperate in a social dilemma as long as the other players also cooperate. Consider the above social dilemma, with $n = 2$, called the Prisoner's Dilemma. In this game, let $CC$ stand for "both players cooperate," let $DD$ stand for "both players defect," let $CD$ stand for "Player 1 cooperates but his partner defects," and let $DC$ stand for "Player 1 defects and his partner cooperates." A self-regarding Player 1 will prefer $DC$ to $CC$, will prefer $CC$ to $DD$, and will prefer $DD$ to $CD$, while an altruistic cooperator will prefer $CC$ to $DC$, will prefer $DC$ to $DD$, and will prefer $DD$ to $CD$; i.e. the self-regarding individual prefers to defect no matter what his partner does, whereas the conditional altruistic cooperator prefers to cooperate so long as his partner cooperates.

Kiyonari et al. (2000) ran an experiment based on this game with real monetary payoffs using 149 Japanese university students. The experimenters ran three distinct treatments, with about equal numbers of subjects in each treatment. The first treatment was a standard "simultaneous" Prisoner's Dilemma, the second was a "second-player" situation in which the subject was told that the first player in the Prisoner's Dilemma had already chosen to cooperate, and the third was a "first-player" treatment in which the subject was told that his decision to cooperate or defect would be made known to the second player before the latter made his own choice. The experimenters found that 38% of the subjects cooperated in the simultaneous treatment, 62% cooperated in the second player treatment, and 59% cooperated in the first-player treatment. The decision to cooperate in each treatment cost the subject about $5 (600 yen). This shows unambiguously that a majority of subjects were conditional altruistic cooperators (62%). Almost as many were not only cooperators, but were also willing to bet that their partners would be (59%), provided the latter were assured of not being defected upon, although under standard conditions, without this assurance, only 38% would in fact cooperate.

### 2.2 Altruistic Punishment

*Strong reciprocity* is an altruistic behavioral propensity often exhibited in daily life and in the laboratory as well. A strong reciprocator is a conditional altruistic cooperator who is willing to punish non-cooperators even when this is personally costly and is unlikely to be compensated by higher material returns in the future. The simplest game exhibiting the altruistic punishment of the strong reciprocator is the *Ultimatum Game* (Güth et al. 1982). Under conditions of anonymity, two players are shown a sum of money, say $10. One of the players, called the Proposer, is instructed to offer any number of dollars, from $1 to $10, to the second player, who is called the Responder. The Proposer can make only one offer and the Responder can either accept or reject this offer. If the Responder accepts the offer, the money is shared accordingly. If the Responder rejects the offer, both players receive nothing. The two players do not face each other again.

There is only *one* Responder strategy that is a best response for a self-regarding individual: accept anything you are offered. Knowing this, a self-regarding Proposer who believes he faces a self-regarding Responder offers the minimum possible amount, $1, and this is accepted.

However, when actually played, the self-regarding outcome is almost never attained or even approximated. In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, Proposers routinely offer Responders very substantial amounts (50% of the total generally being the modal offer) and Responders frequently reject offers below 30% (Güth and Tietz 1990, Camerer and Thaler 1995). Are these results culturally dependent? Do they have a strong genetic component or do all successful cultures transmit similar values of reciprocity to individuals? Roth et al. (1991) conducted the Ultimatum Game in four different countries (United States, Yugoslavia, Japan, and Israel) and found that while the level of offers differed a small but significant amount in different countries, the probability of an offer being rejected did not. This indicates that both Proposers and Responders share the same notion of what is considered fair in that society and that Proposers adjust their offers to reflect this common notion. The differences in level of offers across countries, by the way, were relatively small. When a much greater degree of cultural diversity is studied, however, large differences in behavior are found, reflecting different standards of what it means to be fair in different types of societies (Henrich et al. 2004).

Behavior in the Ultimatum Game thus conforms to the strong reciprocity model: fair behavior in the Ultimatum Game for college students is a 50–50 split. Responders reject offers under 40% as a form of altruistic punishment of the norm-violating Proposer. Proposers offer 50% because they are altruistic cooperators, or 40% because they fear rejection. To support this interpretation, we note that if the offers in an Ultimatum Game are generated by a computer rather than by the Proposer, and if Responders know this, low offers are rarely rejected (Blount 1995). This suggests that players are motivated by *reciprocity*, reacting to a violation of behavioral norms (Greenberg and Frisch 1972). Moreover, in a variant of the game in which a Responder rejection leads to the Responder getting nothing but allows the Proposer to keep the share he suggested for himself, Responders never reject offers, and proposers make considerably smaller (but still positive) offers (Bolton and Zwick 1995). As a final indication that strong reciprocity motives are operative in this game, after the game is over, when asked why they offered more than the lowest possible amount, Proposers commonly said that they were afraid that Responders will consider low offers unfair and reject them. When Responders rejected offers, they usually claimed they want to punish unfair behavior. In all of the above experiments a significant fraction of subjects (about a quarter, typically) conformed to self-regarding preferences.

We should note that while strong reciprocity is a form of moral behavior, it is not generally backed by a high level of moral cognition. Rather, individuals cooperate because it pleases them to do so, and they punish non-cooperators out of anger or pique, not because it is a moral duty to do so. Thus Sanfey et al. (2003) subjected players of the Ultimatum Game to fMRI brain scans, and found that when Responders rejected unfair offers, they exhibited activity in the anterior insula, which is usually associated with emotional responses.

### 2.3 Strong Reciprocity in the Labor Market

Fehr et al. (1997) (see also Fehr and Gächter 1998) performed an experiment to validate what is know as a *gift exchange* model of the labor market. The experimenters divided a group of 141 subjects into "employers" and "employees." The rules of the game are as follows. If an employer hires an employee who provides effort $e$ and receives a wage $w$, his profit is $\pi = 100e - w$. The wage must be between 1 and 100, and the effort is between 0.1 and 1. The payoff to the employee is then $u = w - c(e)$, where $c(e)$ a cost of effort function such that $c(0.1) = 0$, $c(1.0) = 20$. All payoffs involve real money that the subjects are paid at the end of the experimental session. We call this the *Experimental Labor Market Game*.

The sequence of actions is as follows. The employer first offers a "contract" specifying a wage $w$ and a desired amount of effort $e^*$. A contract is made with the first employee who agrees to these terms. An employer can make a contract $(w, e^*)$ with at most one employee. The employee who agrees to these terms receives the wage $w$ and supplies an effort level $e$ that *need not equal the contracted effort* $e^*$. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level, $e \in [0.1, 1]$, with impunity. Although subjects may play this game several times with different partners, each employer-employee interaction is a one-shot (nonrepeated) event. Moreover, the identity of the interacting partners is never revealed.

If employees are self-regarding, they will choose the zero-cost effort level, $e = 0.1$, no matter what wage is offered them. Knowing this, employers will never pay more than the minimum necessary to get the employee to accept a contract, which is 1 (assuming only integer wage offers are permitted). The employee will accept this offer and will set $e = 0.1$. Because $c(0.1) = 0$, the employee's payoff is $u = 1$. The employer's payoff is $\pi = 0.1 \times 100 - 1 = 9$.

In fact, however, this self-regarding outcome rarely occurred in this experiment. The average net payoff to employees was $u = 35$, and the more generous the employer's wage offer to the employee, the higher the effort provided. In effect, employers presumed the strong reciprocity predispositions of the employees, making quite generous wage offers and receiving higher effort, as a means to increase both their own and the employee's payoff, as depicted in figure 1. Similar results have been observed in Fehr, Kirchsteiger, and Riedl ((1993), (1998)).

Figure 1 also shows that, though most employees are strong reciprocators, at any wage rate there still is a significant gap between the amount of effort agreed upon and the amount actually delivered. This is not because there are a few "bad

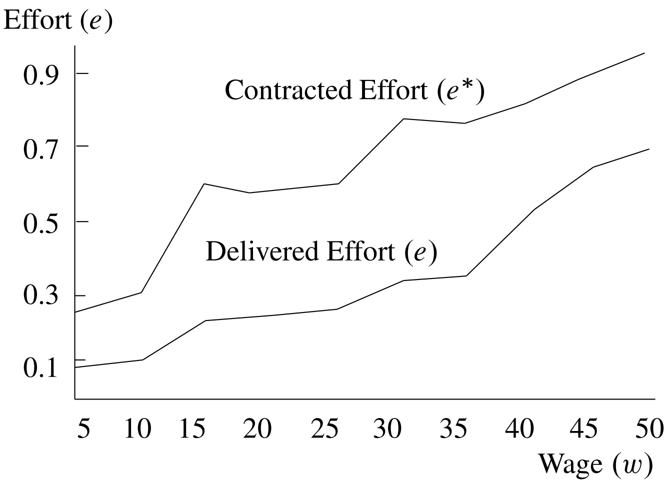

**Figure 1:** Relation of contracted and delivered effort to worker wage (141 subjects). From Fehr, Gächter, and Kirchsteiger (1997).

apples" among the set of employees but because only 26% of the employees delivered the level of effort they promised! We conclude that strong reciprocators are inclined to compromise their morality to some extent.

To see if employers are also strong reciprocators, the authors extended the game by allowing the employers to respond reciprocally to the *actual effort choices* of their workers. At a cost of 1, an employer could *increase* or *decrease* his employee's payoff by 2.5. If employers were self-regarding, they would of course do neither because they would not (knowingly) interact with the same worker a second time. However, 68% of the time, employers punished employees who did not fulfill their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Employers rewarded 41% of employees who *exactly* fulfilled their contracts. Moreover, employees *expected* this behavior on the part of their employers, as shown by the fact that their effort levels *increased significantly* when their bosses gained the power to punish and reward them. Underfulfilling contracts dropped from 71% to 26% of the exchanges, and overfulfilled contracts rose from 3% to 38% of the total. Finally, allowing employers to reward and punish led to a 40% increase in the net payoffs to all subjects, even when the payoff reductions resulting from employer punishment of employees are taken into account.

We conclude from this study that subjects who assume the role of employee conform to internalized standards of reciprocity even when they are certain there are no material repercussions from behaving in a self-regarding manner. Moreover, subjects who assume the role of employer expect this behavior and are rewarded for acting accordingly. Finally, employers reward good behavior and punish bad behavior when they are allowed, and employees expect this behavior and adjust their own effort levels accordingly. In general, then, subjects follow an internalized norm not because it is prudent or useful to do so, or because they will suffer some material loss if they do not, but rather because they desire to do this *for its own sake*.

## 2.4 Altruism and Cooperation in Groups

The *Public Goods Game*, an $n$-person social dilemma, captures many areas of altruistic cooperation in social life, including voluntary contribution to team and community goals. Researchers (Ledyard 1995, Yamagishi 1986, Ostrom et al. 1992, Gächter and Fehr 1999) uniformly find that groups exhibit a much higher rate of cooperation than can be expected assuming the standard model of the self-regarding actor.

A typical Public Goods Game consists of a number of rounds, say 10. In each round, each subject is grouped with several other subjects—say 3 others. Each subject is then given a certain number of points, say 20, redeemable at the end of the experimental session for real money. Each subject then places some fraction of his points in a "common account" and the remainder in the subject's "private account." The experimenter then tells the subjects how many points were contributed to the common account and adds to the private account of *each* subject some fraction, say 40%, of the total amount in the common account. So if a subject contributes his whole 20 points to the common account, each of the 4 group members will receive 8 points at the end of the round. In effect, by putting the whole endowment into the common account, a player loses 12 points but the other 3 group members gain in total 24 (8 times 3) points. The players keep whatever is in their private accounts at the end of the round.

A self-regarding player contributes nothing to the common account. However, most of the subjects do not in fact conform to the self-regarding model. Subjects begin by contributing on average about half of their endowments to the public account. The level of contributions decays over the course of the 10 rounds until in the final rounds most players are behaving in a self-regarding manner. This is, of course, exactly what is predicted by the strong reciprocity model. Because they are altruistic contributors, strong reciprocators start out by contributing to the common pool, but in response to the norm violation of the self-regarding types, they begin to refrain from contributing themselves.

How do we know that the decay of cooperation in the Public Goods Game is due to cooperators punishing free riders by refusing to contribute themselves?

Subjects often report this behavior retrospectively. More compelling, however, is the fact that when subjects are given a more constructive way of punishing defectors, they use it in a way that helps sustain cooperation (Orbell, Dawes, and Van de Kragt 1986, Sato 1987, Yamagishi 1988a,1988b, 1992)

Fehr and Gächter (2000), for instance, used 6- and 10-round Public Goods Games with groups of size 4, and with costly punishment allowed at the end of each round, employing three different methods of assigning members to groups. There were sufficient subjects to run between 10 and 18 groups simultaneously. Under the Partner treatment, the four subjects remained in the same group for all 10 periods. Under the Stranger treatment, the subjects were randomly reassigned after each round. Finally, under the Perfect Stranger treatment, the subjects were randomly reassigned but assured that they would never meet the same subject more than once.

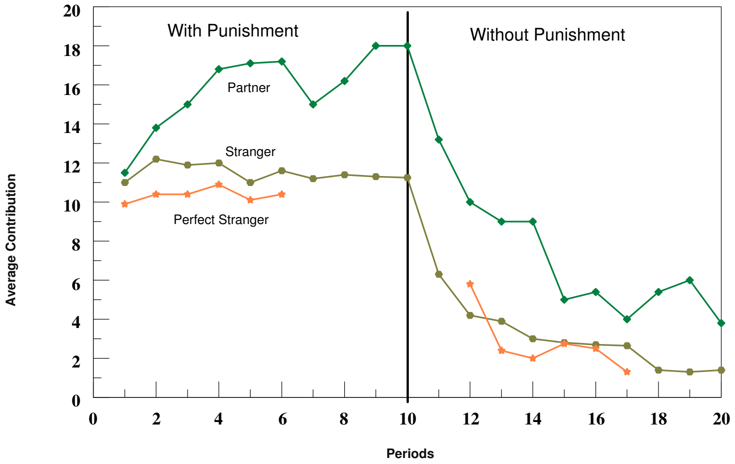

**Figure 2:** Average contributions over time in the Partner, Stranger, and Perfect Stranger Treatments when the punishment condition is played first (Fehr and Gächter 2000).

Fehr and Gächter (2000) performed their experiment for 10 rounds with punishment and 10 rounds without. Their results are illustrated in figure 2. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the Partner game, despite strict anonymity, cooperation increases almost to full cooperation even in the final round. When punishment is not permitted, however, the same subjects experienced the deterioration of cooperation found in previous Public Goods Games. The contrast in cooperation rates between the Partner treatment and the two Stranger treatments is worth noting because the strength of punishment is roughly the same across all treatments. This suggests that the credibility of the punishment threat is greater in the Partner treatment because in this treatment the punished subjects are certain that, once they have been punished in previous rounds, the punishing subjects are in their group. The prosociality impact of strong reciprocity on cooperation is thus more strongly manifested, the more coherent and permanent the group in question.

### 2.5 Character Virtues

*Character virtues* are ethically desirable behavioral regularities that individuals value for their own sake, while having the property of facilitating cooperation and enhancing social efficiency. Character virtues include *honesty*, *loyalty*, *trustworthiness*, *promise keeping*, and *fairness*. Unlike such other-regarding preferences as strong reciprocity and empathy, these character virtues operate without concern for the individuals with whom one interacts. An individual is honest in his transactions because this is a desired state of being, not because he has any particular regard for those with whom he transacts. Of course, the sociopath *Homo economicus* is honest only when it serves his material interests to be so, whereas the rest of us are at times honest even when it is costly to be so and even when no one but us could possibly detect a breach.

Common sense, as well as the experiments described below, indicate that honesty, fairness, and promise-keeping are not absolutes. If the cost of virtue is sufficiently high, and the probability of detection of a breach of virtue is sufficiently small, many individuals will behave dishonestly. When one is aware that others are unvirtuous in a particular region of their lives (e.g., marriage, tax paying, obeying traffic rules, accepting bribes), one is more likely to allow one's own virtue to lapse. Finally, the more easily one can delude oneself into inaccurately classifying an unvirtuous act as virtuous, the more likely one is to allow oneself to carry out such an act.

One might be tempted to model honesty and other character virtues as *self-constituted constraints* on one's set of available actions in a game, but a more fruitful approach is to include the state of being virtuous in a certain way as an argument in one's preference function, to be traded off against other valuable objects of desire and personal goals. In this respect, character virtues are in the same category as ethical and religious preferences and are often considered subcategories of the latter.

Numerous experiments indicate that most subjects are willing to sacrifice material rewards to maintain a virtuous character even under conditions of anonymity. Sally (1995) undertook a meta-analysis of 137 experimental treatments, finding that face-to-face communication, in which subjects are capable of making verbal agreements and promises, was the strongest predictor of cooperation. Of course, face-to-face interaction violates anonymity and has other effects besides the ability to make promises. However, both Bochet et al. (2006) and Brosig et al. (2003) report that only the ability to exchange verbal information accounts for the increased cooperation.

A particularly clear example of such behavior is reported by Gneezy (2005), who studied 450 undergraduate participants paired off to play three games of the following form, all payoffs to which were of the form $(b, a)$, where player 1, Bob, receives $b$ and player 2, Alice, receives $a$. In all games, Bob was shown two pairs of payoffs, $A$:$(x, y)$ and $B$:$(z, w)$ where $x$, $y$, $z$, and $w$ are amounts of money with $x < z$ and $y > w$, so in all cases $B$ is better for Bob and $A$ is better for Alice. Bob could then say to Alice, who could not see the amounts of money, either "Option $A$ will earn you more money than option $B$," or "Option $B$ will earn you more money than option $A$." The first game was $A$:(5,6) vs. $B$:(6,5) so Bob could gain 1 by lying and being believed while imposing a cost of 1 on Alice. The second game was $A$:(5,15) vs. $B$:(6,5), so Bob could gain 1 by lying and being believed, while still imposing a cost of 10 on Alice. The third game was $A$:(5,15) vs. $B$:(15,5), so Bob could gain 10 by lying and being believed, while imposing a cost of 10 on Alice.

Before starting play, Gneezy asked the various Bobs whether they expected their advice to be followed. He induced honest responses by promising to reward subjects whose guesses were correct. He found that 82% of Bobs expected their advice to be followed (the actual number was 78%). It follows from the Bobs' expectations that if they were self-regarding, they would always lie and recommend $B$ to Alice.

The experimenters found that, in game 2, where lying was very costly to Alice and the gain from lying was small for Bob, only 17% of Bobs lied. In game 1, where the cost of lying to Alice was only 1 but the gain to Bob was the same as in game 2, 36% of Bobs lied. In other words, Bobs were loathe to lie but considerably more so when it was costly to Alices. In game 3, where the gain from lying was large for Bob and equal to the loss to Alice, fully 52% of Bobs lied. This shows that many subjects are willing to sacrifice material gain to avoid lying in a one-shot anonymous interaction, their willingness to lie increasing with an increased cost to them of truth telling, and decreasing with an increased cost to their partners of being deceived. Similar results were found by Boles et al. (2000) and Charness and Dufwenberg (2004). Gunnthorsdottir et al. (2002) and Burks et al. (2003) have shown that a socio-psychological measure of "Machiavellianism" predicts which subjects are likely to be trustworthy and trusting.

### 2.6 Norms of Cooperation: Cross-Cultural Variation

Experimental results in the laboratory would not be very interesting if they did not aid us in understanding and modeling real-life behavior. There are strong and consistent indications that the external validity of experimental results is high. In one very important study, Herrmann et al. (2008) had subjects play the Public Goods Game with punishment with 16 subject pools in 15 different countries with highly varying social characteristics (one country, Switzerland, was represented by two subject pools, one in Zurich and one in St. Gallen). To minimize the social diversity among subject pools, they used university students in each country. The phenomenon they aimed to study was *antisocial punishment*.

The phenomenon itself was first noted by Cinyabuguma et al. (2006), who found that some free riders, when punished, responded not by increasing their contributions, but rather by punishing the high contributors! The ostensible explanation of this perverse behavior is that some free riders believe it is their personal right to free-ride if they so desire, and they respond to the "bullies" who punish them in a strongly reciprocal manner—they retaliate against their persecutors. The result, of course, is a sharp decline in the level of cooperation for the whole group.

This behavior was later reported by Denant-Boemont et al. (2007) and Nikiforakis (2008), but because of its breadth, the Herrmann, Thöni, and Gächter study is distinctive for its implications for social theory. They found that in some countries, antisocial punishment was very rare, while in others it was quite common. As can be seen in figure 3, there is a strong negative correlation between the amount of anti-punishment exhibited and the World Democracy Audit's assessment of the level of democratic development of the society involved.

Figure 4 shows that a high level of antisocial punishment in a group translates into a low level of overall cooperation. The researchers first ran 10 rounds of the Public Goods Game without punishment (the $N$ condition), and then another 10 rounds with punishment (the $P$ condition). The figures show clearly that the more democratic countries enjoy a higher average payoff from payoffs in the Public Goods Game.

How might we explain this highly contrasting social behavior in university students in democratic societies with advanced market economies on the one hand,

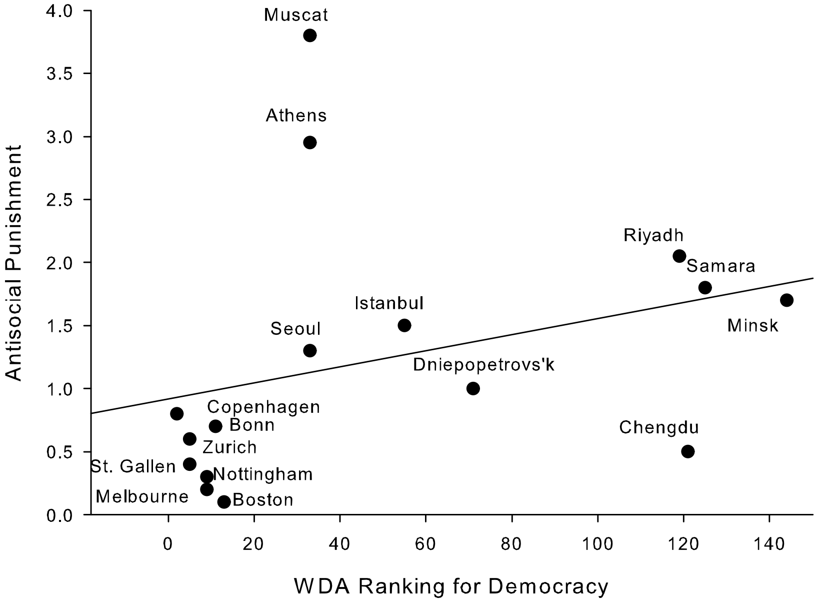

**Figure 3:** Countries judged highly democratic (political rights, civil liberties, press freedom, low corruption) by the World Democracy Audit engage in very little antisocial punishment, and conversely. (Statistics from Herrmann, Thöni, and Gächter, 2008.)

and more traditional societies based on authoritarian and parochial social institutions on the other? The success of democratic market societies may depend critically upon moral virtues as well as material interests, so the depiction of economic actors as *Homo economicus* is as incorrect in real life as it is in the laboratory. These results indicate that individuals in modern democratic capitalist societies have a deep reservoir of public sentiment that can be exhibited even in the most impersonal interactions with unrelated others. This reservoir of moral predispositions is based upon an innate prosociality that is a product of our evolution as a species, as well as the uniquely human capacity to internalize norms of social behavior. Both forces predispose individuals to behave morally, even when this conflicts with their material interests, and to react to public disapprobation for free-riding with shame and penitence rather than antisocial self-aggrandizement.

More pertinent to the purposes of behavioral game theory, this experiment shows that laboratory games can be deployed to shed light on real-life social regularities that cannot be explained by participant observation or cross-country statistical analysis alone.

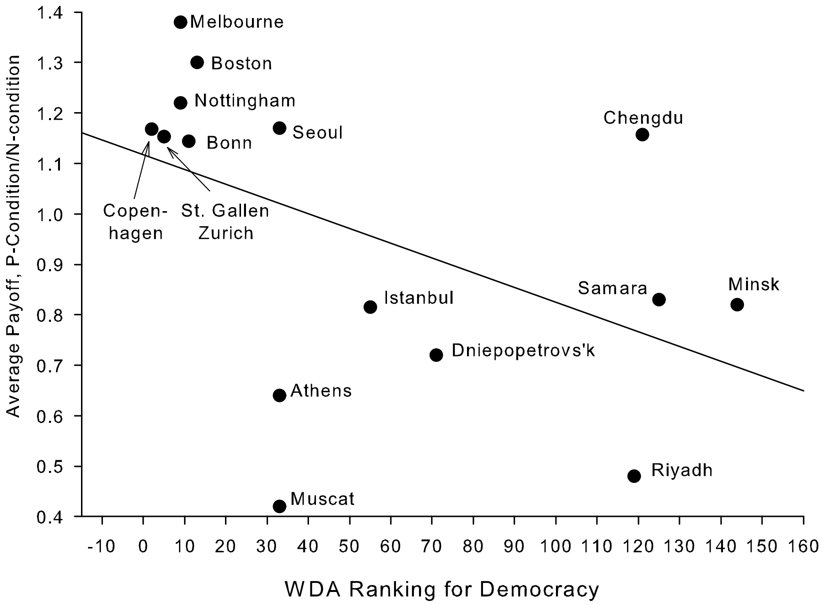

**Figure 4:** Antisocial punishment leads to low payoffs (Statistics from Herrmann, Thöni, and Gächter, Online Supplementary Material, 2008).

## 3 Conclusion

The problems behavioral ethics poses for moral philosophy include the finding that individuals tend to treat morality the same way they treat a standard consumer good: the higher the price of behaving morally, the less likely they are to do so. Moreover, subjects exhibit forms of moral behavior, including strong reciprocity, which is not considered moral by most contemporary ethical theories. It can be argued that popular views of ethics are no more relevant to modern moral philosophy than popular accounts of physical phenomena are to modern physics. If, for instance, we discover that cannibalism is morally prohibited, the fact that humans have practiced cannibalism for tens of thousands of years is no more a problem than our believing the big bang theory of the origins of the universe, despite the fact this theory corresponds to the origins story of any known society. On these grounds, universalist moral theories, including utilitarian and deontological, can simply ignore popular morality as a source of insight, and can reject the notion that observed human behavior is the ultimate arbiter among ethical theories.

While I believe physicists are correct in rejecting folk physics as a basis for adjudicating among alternative approaches to the laws of physics, the same cannot be said of philosophers who reject behavioral ethics. While physicists have a proven track record of developing ever more powerful models of physical reality, philosophers have only a long history of alternative models of morality none of which explains anything about the material world, and none of which is embraced by more than a few experts in the field. Aristotle and St. Thomas Aquinas are read today not just as historical *reliquia*, but as living documents from which new insights can be drawn. Even within a contemporary philosophical tradition it is rare for two philosophers to agree on all major issues. Moral philosophy is thus more like art or literature in creatively illuminating the human condition without generating the sort of demonstrable truths that would be needed to justify ignoring actual human moral behavior.

If morality is identified with moral behavior, the most salient and immediate implication is that moral truths are relative to particular cultures: individuals can disagree on the content of morality in a manner that cannot be adjudicated through a careful examination of the evidence or because the individuals differ concerning matters of fact that, if resolved, would lead their moral differences to evaporate. Moral relativism of the sort suggested by behavioral ethics is widely rejected by moral philosophers because moral relativism appears not to leave a role for a reasoned investigation of morality at all. *De gustibus*, as the saying goes, *non est disputandum*. More important, if we relinquish the notion of a universal morality, then must we not accept a situation in which there is no real right or wrong, but only differences in what people believe to be right or wrong? Does it not follow from this that since our moral beliefs have no status privileged by our superior expertise, education, or scholarly dedication, are we not obliged to tolerate moral beliefs and practices that we consider vile, abhorrent, and disgusting?

The answer philosopher David Wong gives to these questions in Nature Moralities (2006) is in the negative. Wong's alternative to universality, which he calls pluralistic relativism, holds that "there is no single true morality. However, it recognizes significant limits on what can count as a true morality." (p. xii) The reason for these limitations is that morality is the product of the evolutionary history of our species, serving the role of social cohesion by endowing all, or at least most, members of a group whose survival depends on cooperation with a set of common commitments, expectations, and conventions that promote group solidarity. Wong's explanation is clearly dependent upon the facts of human existence, and is far from the sort of empirical blindness that is favored by many philosophers.

I suspect there will be a bright future for pluralistic relativism in meta-ethics. The sort of virtue theory first proposed by Aristotle and revived in recent years by G. E. M Anscombe, Philippa Foot, Martha Nussbaum and Amartya Sen interacts fruitfully with pluralistic relativism, and thus also complements the scientific findings of behavioral ethics

## References

Andreoni, James and John H. Miller, "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica* 70,2 (2002):737–753.

Blount, Sally, "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior & Human Decision Processes* 63,2 (August 1995):131–144.

Bochet, Olivier, Talbot Page, and Louis Putterman, "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior and Organization* 60,1 (2006):11–26.

Boles, Terry L., Rachel T. A. Croson, and J. Keith Murnighan, "Deception and Retribution in Repeated Ultimatum Bargaining," *Organizational Behavior and Human Decision Processes* 83,2 (2000):235–259.

Bolton, Gary E. and Rami Zwick, "Anonymity versus Punishment in Ultimatum Games," *Games and Economic Behavior* 10 (1995):95–121.

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

Brosig, J., A. Ockenfels, and J. Weimann, "The Effect of Communication Media on Cooperation," *German Economic Review* 4 (2003):217–242.

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

Burks, Stephen V., Jeffrey P. Carpenter, and Eric Verhoogen, "Playing Both Roles in the Trust Game," *Journal of Economic Behavior and Organization* 51 (2003):195–216.

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

Cavalli-Sforza, Luca L. and Marcus W. Feldman, "Theory and Observation in Cultural Transmission," *Science* 218 (1982):19–27.

Charness, Gary and Martin Dufwenberg, "Promises and Partnership," October 2004. University of California at Santa Barbara.

Cinyabuguma, Matthias, Talbott Page, and Louis Putterman, "Can Second-Order Punishment Deter Perverse Punishment?," *Experimental Economics* 9 (2006):265–279.

Denant-Boemont, Laurent, David Masclet, and Charles Noussair, "Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment," *Economic Theory* 33,1 (October 2007):145–167.

Dunbar, R. I. M., “Coevolution of Neocortical Size, Group Size and Language in Humans,” *Behavioral and Brain Sciences* 16,4 (1993):681–735.

Fehr, Ernst and Simon Gächter, “How Effective Are Trust- and Reciprocity-Based Incentives?,” in Louis Putterman and Avner Ben-Ner (eds.) *Economics, Values and Organizations* (New York: Cambridge University Press, 1998) pp. 337–363.

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

— , Georg Kirchsteiger, and Arno Riedl, “Does Fairness Prevent Market Clearing?,” *Quarterly Journal of Economics* 108,2 (1993):437–459.

— , — , and — , “Gift Exchange and Reciprocity in Competitive Experimental Markets,” *European Economic Review* 42,1 (1998):1–34.

— , Simon Gächter, and Georg Kirchsteiger, “Reciprocity as a Contract Enforcement Device: Experimental Evidence,” *Econometrica* 65,4 (July 1997):833–860.

Gächter, Simon and Ernst Fehr, “Collective Action as a Social Exchange,” *Journal of Economic Behavior and Organization* 39,4 (July 1999):341–369.

Gneezy, Uri, “Deception: The Role of Consequences,” *American Economic Review* 95,1 (March 2005):384–394.

Greenberg, M. S. and D. M. Frisch, “Effect of Intentionality on Willingness to Reciprocate a Favor,” *Journal of Experimental Social Psychology* 8 (1972):99–111.

Gunnthorsdottir, Anna, Kevin McCabe, and Vernon Smith, “Using the Machiavellianism Instrument to Predict Trustworthiness in a Bargaining Game,” *Journal of Economic Psychology* 23 (2002):49–66.

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

— , R. Schmittberger, and B. Schwarze, “An Experimental Analysis of Ultimatum Bargaining,” *Journal of Economic Behavior and Organization* 3 (May 1982):367–388.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies* (Oxford: Oxford University Press, 2004).

Herrmann, Benedikt, Christian Thöni, and Simon Gächter, “Anti-social Punishment Across Societies,” *Science* 319 (7 March 2008):1362–1367.

Kiyonari, Toko, Shigehito Tanida, and Toshio Yamagishi, “Social Exchange and Reciprocity: Confusion or a Heuristic?,” *Evolution and Human Behavior* 21

(2000):411–427.

Ledyard, J. O., “Public Goods: A Survey of Experimental Research,” in John H. Kagel and Alvin E. Roth (eds.) *The Handbook of Experimental Economics* (Princeton: Princeton University Press, 1995) pp. 111–194.

Nikiforakis, Nikos S., “Punishment and Counter-punishment in Public Goods Games: Can we Still Govern Ourselves?,” *Journal of Public Economics* 92,1–2 (2008):91–112.

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

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

Richerson, Peter J. and Robert Boyd, *Not By Genes Alone* (Chicago: University of Chicago Press, 2004).

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, “Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study,” *American Economic Review* 81,5 (December 1991):1068–1095.

Sally, David, “Conversation and Cooperation in Social Dilemmas,” *Rationality and Society* 7,1 (January 1995):58–92.

Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen, “The Neural Basis of Economic Decision-Making in the Ultimatum Game,” *Science* 300 (13 June 2003):1755–1758.

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

Varian, Hal R., “The Nonparametric Approach to Demand Analysis,” *Econometrica* 50 (1982):945–972.

Wong, David B., *Natural Moralities: A Defense of Pluralistic Relativism* (Oxford: Oxford University Press, 2006).

Yamagishi, Toshio, “The Provision of a Sanctioning System as a Public Good,” *Journal of Personality and Social Psychology* 51 (1986):110–116.

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

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

— , “Group Size and the Provision of a Sanctioning System in a Social Dilemma,” in W. B. G. Liebrand, David M. Messick, and H. A. M. Wilke (eds.) *Social*

*Dilemmas: Theoretical Issues and Research Findings* (Oxford: Pergamon Press, 1992) pp. 267–287.

c\Papers\behavioral and evolutionary ethics\BehavioralEthics.tex September 6, 2009