politics, philosophy & economics article

London
Thousand Oaks, CA
and New Delhi

1470-594X
200602 5(1) 5–32

Behavioral ethics meets natural justice

Herbert Gintis
Santa Fe Institute, USA and Central European University, Hungary

abstract Binmore's *Natural Justice* offers an evolutionary approach to morality, in which moral rules form a cultural system that is robust and evolutionarily stable. The folk theorem is the analytical basis for his theory of justice. I argue that this is a mistake, as the equilibria described by the folk theorem lack dynamic stability in games with several players. While the dependence of Binmore's argument on the folk theorem is more tactical than strategic, this choice does have policy implications. I do not believe that moral rules are solutions to the Nash bargaining problem. Rather, I believe that human beings are emotionally constituted, by virtue of their evolutionary history, to embrace prosocial and altruistic notions of in-group/out-group identification and reciprocity. These aspects of human nature are incompatible with Binmore's notion that humans are self-regarding creatures. I present empirical evidence supporting a specific form of human, other-regarding preferences known as strong reciprocity.

keywords justice, ethics, folk theorem, evolutionary game theory

1. Introduction

Ken Binmore is at once a first-rate scientist and mathematician, an exemplary philosopher, and a mind well read in the behavioral disciplines. He is also firmly grounded in the analytical intellectual tradition of Locke and Hume, and has little tolerance for the obfuscation of the Continental tradition exemplified by Immanuel Kant.

I like that. Indeed, the leading quote in my recent game theory book[1] was Ludwig Wittgenstein's famous 'Was sich sagen läßt, läßt sich klar sagen, und wovon man nicht sprechen kann, darüber muß man schweigen.'[2] Another of Ken's admirable traits is his willingness, indeed, predilection to 'shoot from the

DOI: 10.1177/1470594X06060617
Herbert Gintis, 15 Forbes Avenue, Northampton, MA 01060, USA [email: hgintis@comcast.net]

hip' rather than proffer tepidly guarded positions. I generally agree with his conclusions, including his impatience with the dominant tradition in philosophical ethics, which holds that ethical principles can be deduced from reason alone,[3] and the short shrift he gives to the Kantian tradition which, for instance, concludes that cooperating in the prisoner's dilemma is a deontological imperative,[4] and his proudly wearing the mantle of 'sociobiology', which many refuse to don because of the contumely (unfairly) heaped upon the term in the years following Edward O. Wilson's pathbreaking book.[5]

Binmore's new book, *Natural Justice*, offers us an evolutionary approach to morality, in which moral rules form a cultural system that grew historically with the emergence of Homo sapiens and is evolutionarily robust and stable. 'All the societies studied by anthropologists that survived into modern times with a pure hunter-gathering economy', Binmore accurately notes, 'had similar social contracts with a similar deep structure . . . They tolerate no bosses, and they share on a very egalitarian basis.' Such rules must be efficient in the sense that societies that use these rules prosper at the expense of those that do not, and they must be fair or they will not be accepted by those whom such rules govern.[6] Binmore is well aware of the various weaknesses of the Nash equilibrium concept and its static refinements.[7] He nevertheless accepts the folk theorem from game theory as the analytical basis for a theory of justice, despite its dependence on static Nash equilibrium criteria. I shall argue that this ~~also~~ is a mistake, as the equilibria described by the folk theorem virtually never have attractive dynamic stability properties in games with several players. The issue of dynamic versus static equilibrium criteria is, in principle, quite simple. A Nash equilibrium is a choice of a strategy by each player, such that given the choices of the other players, no player can gain by altering his own strategy. Dynamic stability requires that when all strategies are perturbed *at the same time*, there is a long-run tendency for the players' choices to return to the equilibrium. The dynamic criterion is much stronger that the Nash criterion, and even such common refinements as sub-game perfection (a Nash equilibrium remains Nash in every sub-game), trembling-hand perfection (if all players make errors with low probability, the resulting equilibrium tends toward no-error equilibrium as the error rate tends to zero), or sequential equilibrium (players use best responses at all nodes, on and off the equilibrium path) generally do not ensure, or even render likely, dynamic stability.

I suspect that the dependence of Binmore's argument on the folk theorem is more tactical than strategic. Binmore certainly does not need the full force of the folk theorem. At most, he requires that the portion of the Pareto frontier that is Pareto superior to the mutual defect equilibrium be accessible. Probably, it would be sufficient to argue that society reaches a Nash equilibrium with a positive level of cooperation, and that this equilibrium has plausible dynamic stability properties.

My disagreement with Binmore on this count does, however, impact upon the

social policy implications of ethical theory. I agree with Binmore that one of the least attractive aspects of traditional philosophical ethicists is their tendency to 'try to force their aspirations on others by appealing to some invented source of absolute authority'.[8] Even such great philosophers as Hobbes, Locke, Rawls, and Nozick are guilty of such hubris. Like Binmore, I believe moral principles are facts in the world, and the evolution and transformation of ethical principles follow natural laws which, if we understand them, can be successfully altered to improve the lives of people. However, by contrast with Binmore, I do not believe these moral rules involve applying local cultural-social indices to the Nash bargaining problem, using the concept of the original position espoused by Rawls and Harsanyi. Rather, I believe that human beings are emotionally constituted, by virtue of their evolutionary history, to embrace prosocial and altruistic notions of in-group/out-group identification and reciprocity. These aspects of human nature are incompatible with Binmore's notion that humans are self-regarding creatures.[9]

Binmore implicitly identifies 'rationality' with having self-regarding preferences, and considers actions that violate the principle of self-regard as 'irrational'. However, experimental evidence supports the notion that human beings have preferences that are *other regarding* in the sense that people care about payoffs to others, not just themselves, and individuals care about how outcomes are generated, not just the outcomes themselves. I am thus one of the 'behavioral economists' whom Binmore critiques in the following terms: 'There is a school of behavioral economists who seem to believe that real people always behave as though maximizing some utility function, albeit one that depends on parameters that are commonly neglected in traditional economics . . . I believe that economists and game theorists need to face up to the fact that human behavior is often downright irrational.'[10]

Binmore holds that 'the problem isn't that boundedly rational people maximize something unusual, but that they don't maximize anything at all'.[11] I believe this flies directly in the face of the evidence, and conflicts with a correct understanding of the concept of preference functions that underlies the scientific study of behavior. The point, of course, is not whether people always maximize, but whether they maximize sufficiently frequently and their deviations from maximization are sufficiently minor that their behavior can best be described by a model in which they maximize. This I assert is the case.

My remarks expand on these points. First, I shall dispute Binmore's use of the folk theorem. Second, I will argue from evolutionary principles that utility maximization should be a central tool in analyzing human behavior, even if humans are not self-regarding. Third, I will present empirical evidence supporting a specific form of human, other-regarding preferences known as *strong reciprocity*.[12] Fourth, I will provide evidence from the sociological literature to the effect that human preferences are partially 'programmable' through socialization, and the dominant culture in most societies promotes prosocial values.

Fifth, I will analyze charity in our society, showing that it conforms to notions of strong reciprocity rather than fair bargaining. I remind the reader that my objections in no way conflict with the conceptual framework laid down by Binmore in *Natural Justice* and his previous two books.[13]

2. The folk theorem as a model of social cooperation

The folk theorem holds that high levels of cooperation can be attained among self-regarding individuals facing a social dilemma if information is public, signals are highly accurate, and individuals are very patient.[14] But in a large class of interactions, all three assumptions are implausible empirically, and their plausibility declines rapidly with increasing group size. Moreover, even when all these conditions are met, the extant models of cooperation among self-regarding agents provide no reason to believe that high levels of cooperation could have evolved when rare in a population, or that cooperation could be sustained over long periods assuming any plausible evolutionary dynamic. In short, in the relevant settings, highly cooperative equilibria among self-regarding agents are both evolutionarily inaccessible and unstable. Not surprisingly, then, the mechanisms that ensure cooperation in these models are not those generally observed in cooperative groups. Five empirical problems are especially important.[15]

First, if the self-regarding models were correct, it is not clear why humans would ever cooperate or punish in one-shot, anonymous interactions, where the carefully constructed incentives for self-regarding cooperation are conspicuously absent. As I show in Section 4, however, a high level of cooperation can frequently be achieved in such situations. Ken Binmore attributes this behavior to human error, but there are good grounds for rejecting this interpretation.

Second, the equilibrium concepts in economic models are sub-game perfection and sequential equilibrium. These refinements of the Nash equilibrium are desirable conditions, but they are far from sufficient in a dynamic setting. Real-world social relationships must have evolved historically under adverse and primitive conditions, and must be capable of withstanding invasions by mutant strategies. By contrast, repeated game models with many agents are, in every case I know, dynamically unstable and tend to fail when signals are noisy and private,[16] and it has never been shown that they could be repaired to have the dynamic stability properties that render the evolution of cooperative institutions possible and ensure their structural continuity through time.

The reason such equilibrium concepts are inadequate is straightforward. Subgame perfection and sequential equilibrium require, at every point in the game tree, that no single agent can gain by deviating from the equilibrium strategy. Dynamic stability, by contrast, requires that no *subset* of players can gain from a coordinated deviation from the equilibrium strategy. The latter requirement is usually considerably more demanding than the former. Moreover, even when simultaneous deviations from equilibrium are considered, only the properties of

the model when the errors are infinitesimal are considered. Simulations show that even very small positive errors (less than 1 percent) can compromise these cooperative equilibria.

Two aspects of self-regarding cooperation models make it unlikely that dynamically stable versions could be developed. First, cooperative equilibria are not isolated points. Rather, every neighborhood of an equilibrium contains a distinct equilibrium. This remains the case even if attention is limited to equilibria on the efficiency frontier of feasible payoffs. Since there is a conflict of interest among players as to which equilibrium should be played, there is no mechanism through which a deviation from one equilibrium will lead to a return to that equilibrium. Second, if only pure strategies are observable, or if information is private, then all efficient equilibria involve each agent using mixed strategies (that is, strategies in which agents randomize over their pure strategies). We know that plausible dynamic processes (so-called aggregate monotone dynamics) with more than one population of agents only support strict equilibria, in which agents play only pure strategies.[17] While self-regarding cooperation models generally assume a single population, they all assume agents choose at multiple points in the stage game. Therefore, the equivalent agent extensive form game (in which a player who chooses at distinct information sets is replaced by distinct players with the same payoffs) is necessarily unstable.

Third, the folk theorem shows that cooperation can be sustained if agents are sufficiently patient (that is, the discount factor is sufficiently close to unity). However, individual discount factors are likely to have been low throughout most of human history, both because of the riskiness of life and the fragility of group ties, on the one hand,[18] and the tendency of humans to exhibit high short-term discount rates, on the other.[19] For instance, hunter-gather groups typically experience periodic threats to their existence, in the form of pestilence, famine, and war, at which time the discount factor is quite low, since the probability of group dissolution is high. Self-regarding cooperation models predict the dissolution of such groups, whereas behavioral models predict that such conditions may favor the emergence of agents who cooperate and punish without regard to the discount factor. Experiments in behavioral economics (described in Section 4) show that such agents do exist in large numbers.

Fourth, the folk theorem models that have plausible stability properties are those in which, when shirking is detected, the group reverts to noncooperation for a sufficiently large number of periods that it is not profitable to shirk. These are called 'trigger strategies'.[20] In small groups with highly accurate public signals, trigger strategy models are quite robust. However, in larger groups with private or imperfect signals, these models lead to very low levels of cooperation. It is not surprising, then, that *trigger strategies are rarely observed to be the strategic mechanism through which cooperation is maintained in empirical studies of hunter-gatherer societies or other forms of social cooperation involving more than a few agents*.[21] The inefficiency of trigger strategies is due to the fact that

all group members are punished for the sins of each. More sophisticated folk theorem models are able to target the shirker, the other members acting in concert to impose penalties sufficiently large that shirking is not profitable. This is, of course, quite plausible, as this is the nature of punishment in real groups. In these models, however, since all agents are self-regarding, none will voluntarily punish others if this involves a positive cost to the punishers. In effect, we have a *second-order, free-rider problem*. The self-regarding models must thus deploy a mechanism for punishing those who fail to punish others.[22] As I argue below, these second-order punishment methods have poor stability properties. The fifth empirical problem with such models is the perhaps more important fact that, while there is a high frequency of punishment of norm violators in social groups, *second-order punishment is virtually never observed in social groups of more than a few agents*. An individual who refuses to participate in punishing a malefactor is simply left in peace.[23]

2.1. Cooperation in repeated games with public information

The previous section showed that current models of cooperation in large groups have little or no explanatory power. While it is always possible for someone to discover a satisfactory alternative based on self-regarding agents, the informational imperfections analyzed in this section suggest that this is unlikely to occur. By contrast, I show here that a small amount of other-regarding behavior (in particular, the willingness to punish defectors) dramatically improves the efficiency and stability of models of cooperation, rendering them fully capable of explaining cooperation in large groups even under adverse informational conditions.

Consider a group of size n, in which each member can work or shirk in each time period $t = 1, 2, \ldots$. The cost of working is $c > 0$ and the benefit to the group is $b > c$, shared equally among other group members (note that any share received by the benefactor himself is reflected in a smaller c). Clearly, a self-regarding member will shirk in a one-shot game. However, suppose the game is repeated in each period, and all agents have discount factor δ. The value of working, assuming all other members work, is then $v_c = b - c + \delta v_c$, which gives

$$v_c = \frac{b - c}{1 - \delta} \quad . \tag{1}$$

Suppose that this arrangement continues until a member shirks, upon which the group dissolves and all members receive a payoff of zero in all future periods. To see if this repeated game has full cooperation as a sub-game perfect Nash equilibrium, we must check that no member has an incentive to shirk in any period. The value of shirking when all others work is b, since the shirker receives an amount $b/(n - 1)$ from each of the other $n - 1$ members (we assume the defection is not detected until the end of the period, so the shirker receives an equal share of the total benefit). The condition for cooperation is thus $v_c \geq b$ or

$$\delta \geq \frac{c}{b} . \tag{2}$$

It is unrealistic, however, to assume that there are no errors. Indeed, in everyday life, errors (intended cooperation that appears to others as shirking) perhaps occur with an order of magnitude of 5 percent or even 10 percent. Suppose, for instance, a working agent fails to produce the benefit b, and appears to the other members of the group to be shirking, with a probability $\epsilon > 0$. In Appendix A, I show that as group size becomes large, cooperative efficiency approaches very low levels. Simulations with plausible parameter values show that even for groups of 10–15 members, cooperative efficiency is likely to be low.

The problem with the model is that the only way to punish a defector is to punish every member of the group. Consider the obvious alternative of punishment directed at the offender alone. Suppose a defector receives punishment p from the group, at punishment cost c_p to the group. We assume the costs are shared, so with full cooperation, each member pays $(n-1)\,\epsilon c_p/(n-1) = \epsilon c_p$ per period in punishing others (assuming a defector does not punish himself) and receives punishment ϵp. Since each member supplies benefit $b/(n-1)$ to each of the $n-1$ other members with probability $1-\epsilon$, and no benefit with probability ϵ, each member's expected benefit is $b\,(1-\epsilon)$. So the value of this new game is

$$v_c = b(1-\epsilon) - c - \epsilon(p + c_p) + \delta v_c , \tag{3}$$

which simplifies to

$$v_c = \frac{b(1-\epsilon) - c - \epsilon(p + c_p)}{1-\delta} , \tag{4}$$

The value of defecting for one period and then returning to cooperation is

$$b(1-\epsilon) - p - \epsilon c_p + \delta v_c , \tag{5}$$

assuming the agent continues to punish other observed shirking. It is easy to check that Equation 5 is less than v_c precisely when

$$c < p\,(1-\epsilon) . \tag{6}$$

Note that this is *independent of both group size and the discount factor*, so this solution to the problem of cooperation is extremely attractive. However, it obviously suffers from the second-order, free-rider problem: Why should a self-regarding member punish another member? Of course, we could simply add another layer of costly punishment of non-punishers, but this just pushes the problem to the next level. In our models, by contrast, agents punish without material incentives to do so.

The answer given by Fudenberg and Maskin in their seminal paper on perfect public information models was to enforce costly punishment.[24] Thus, if a member fails to punish, all members punish the non-punisher sufficiently harshly that the non-punisher's gain is wiped out. If there are no errors in commission of the punishment action, this solution is quite effective, as this second-order punishment will never be needed. This approach was extended to imperfect public information models by Abreu, Pearce, and Stacchetti and by Fudenberg, Levine, and Maskin, who showed that close to full cooperation can be attained for sufficiently long-lived agents if the information structure is sufficiently rich to detect shirkers.[25]

However, suppose in our model that there is some error ~~of commission~~ in the agents' act of punishing defectors. For simplicity, suppose this error rate is ϵ, the same as the error rate of cooperation. In Appendix B, I show that cooperation can only be sustained for small-sized groups, using plausible values for ϵ and the discount factor, δ.

2.2. Standing models and private information

In the models discussed up to this point, each member of the group receives the same, perhaps imperfect, signal of the behavior of each other member. In most empirically relevant cases, however, different group members will receive different, and perhaps conflicting, signals concerning other members. For instance, I might see someone sleeping under a tree when he should be hunting, but no other group member may be in the vicinity to witness the scene. To illustrate the problems that arise with private signals, we shall see that a very robust public information self-regarding agent model quickly deteriorates when the information becomes private even to a relatively small degree. Models of cooperation with other-regarding preferences, we should note, assume private information, yet exhibit a high level of cooperation.

Robert Sugden and Robert Boyd have developed a public information model using the notion of group members being either in *good standing* or *bad standing*.[26] Consider the *standing model* version of Equation 4. At the beginning of the game, all agents are categorized as in good standing. At the end of each time period, an agent is in good standing unless he has either defected or failed to punish a member currently in bad standing, in which case he is in bad standing. Let us first consider the public signal version of the model, in which a cooperative signal is incorrectly seen as a defect signal with probability ϵ, and the signal that an individual punished a noncooperator is seen as a failed-to-punish signal with the same probability ϵ. Even with full cooperation, on average $(n-1)\,\epsilon\,(1 + (n-2)\epsilon)$ other agents will signal their having defected, and for each such signaled defection, $(n-2)\epsilon$ other agents will signal their having failed to punish. Therefore, if a member cooperates, he will punish $(n-1)\epsilon(1 + (n-2)\epsilon)$ members, at an expected cost of $c_p(n-1)\epsilon(1 + (n-2)\,\epsilon)/(n-1) = \epsilon c_p(1 + (n-2)\epsilon)$. Moreover, a member will be obliged to cooperate or punish a noncooperator

$1 + (n-1)\epsilon(1 + (n-2)\epsilon)$ times, so will be perceived to have fully cooperated with probability

$$p^* = (1-\epsilon)^{1 + (n-1)\epsilon(1 + (n-2)\epsilon)}. \tag{7}$$

We then have

$$v_c = b(1-\epsilon) - c - p(1-p^*) - \epsilon c_p(1 + (n-2)\epsilon) + \delta v_c. \tag{8}$$

The first two terms on the right-hand side of this equation are the direct benefits and costs of cooperating. The third term is the expected punishment he will receive (since p^* is the probability that he is considered to have cooperated) and the fourth term is the expected cost of punishing others. The final term is the discounted value of returning to cooperation. This expression simplifies to

$$v_c = \frac{b(1-\epsilon) - c - p(1-p^*) - \epsilon c_p(1 + (n-2)\epsilon)}{1-\delta}, \tag{9}$$

and the gain from defecting for one period and then returning to cooperating and punishing is $b(1-\epsilon) - p + \delta v_c$. This gives rise to the following inequality for cooperation:

$$p(1-\epsilon)^{1 + (n-1)\epsilon(1 + (n-2)\epsilon)} > c + c_p\epsilon(1 + (n-2)\epsilon). \tag{10}$$

Using the same parameters as before, we find that cooperation can be sustained even with error rates as high as 13.9 percent, but for error rates in excess of 5.5 percent, the return to cooperation is negative, because there is so much punishment being meted out.

This example vividly illustrates the Achilles heel of standing models of cooperation: even with public signals, the informational requirements are implausible. Each agent must know the standing of $n-1$ other agents and react to the standing of, on average, $(n-1)\epsilon(1 + (n-2)\epsilon)$ agents minus $(n-1)\epsilon$ agents who defected and $(n-1)\epsilon(n-2)\epsilon$ agents who failed to punish a defection. The individual error rate per period is then approximately $(n-1)^2\epsilon^2$. This assumes, in addition, that there are no errors of perception. With errors of perception, some members misread the status of some other members, and mistakenly punish members in good standing. If errors in perception occur at rate ϵ, since each agent observes $n-1$ other agents, each of whom makes $n-1$ punishing/non-punishing decisions, each agent makes an average of approximately $(n-1)^2\epsilon$ perception errors, and hence the error rate per period for the group as a whole is approximately $n(n-1)^2\epsilon$. For even relatively small values of ϵ, say $\epsilon \approx 1/n$, this aggregate error rate may be extremely high.

The informational demands of a private information model are much more modest than that of a public information model. For instance, we could plausibly assume that each member of a group of size n receives information from a

subset of size k, no matter how large n may grow. Informational requirements then grow linearly with n, rather than quadratically, as in the previous example. But how can we ensure defectors are disciplined in a private information world? There is evidence that standing models do capture some real-world, decentralized, cooperation-inducing institutions, but only defections affect status, not failures to punish. In the absence of second-order punishment, punishers must be altruistic in one form or other.

The obvious next step is to consider more general ways to use private information efficiently. This is in fact the tack taken in recent years by several economists.[27] The technical problems involved in developing an equilibrium with a high level of cooperation and assuming private information and self-regarding agents are extreme. If punishing an observed breech of cooperative norms is costly, and if team members generally do not know which members observed which breeches, costly first-order punishment will not occur because those who see the defection know that they will not be punished for failing to punish. Therefore, first-order punishment must fail to be costly. There are various ways of achieving this result involving the use of mixed strategy sequential equilibria, so these models are vulnerable to the critique that the mechanisms involved are not seen empirically and they have very poor stability properties.

I argued above that there is no reason to believe that a sequential Nash equilibrium of a repeated game will have any particularly valuable dynamic stability properties. To illustrate this, I have constructed an agent-based simulation of the Bhaskar-Obara model of cooperation with private information, in the Pascal programming language, as implemented by Borland Delphi 6.0.[28] The stage game is as above, with $b = 3$ and $c = 1$. Agents are randomly assigned to groups of size n in each of 100,000 periods. In each period, each group plays the stage game repeatedly, the game terminating with a probability of 0.05 at the end of each round, and thus implementing a discount factor of 0.95. The simulation begins by creating 210 agents, each endowed at time of creation with two parameters. The first, *DefectRound*, indicates at which round the agent will voluntarily defect. If this is very large, the agent never defects. Since we wish to assess the stability of equilibrium rather than whether it is globally stable, the program initially assigns 80 percent of agents with *DefectRound* = 100, which effectively means they never defect. The other 20 percent of agents are randomly assigned *DefectRound* values between 1 and 10. The second parameter is *Tolerance*, which indicates how many defections an agent who voluntarily cooperates must see before beginning to defect. All agents are assigned *Tolerance* = 0, so they defect at the first defection signal they receive (this is the equilibrium value for the Bhaskar-Obara model).

In each round, for each group, each member sends a signal indicating whether he cooperated or defected, with error rate ϵ, to every other group member. On the basis of this signal, all agents then update their willingness to cooperate in the next round. As soon as the round hits or exceeds an agent's *DefectRound*, or he

accumulates more than *Tolerance* defect signals, the agent defects from that point on with this particular group.

At the end of every 100 periods, the simulation implements a reproduction phase, using the relative fitness of the agents as measured by their accumulated score over the 100 periods, and replacing 5 percent of poorly performing agents by copies of better performing agents. We implement this by a simple imitation process that has the same dynamic properties as the replicator dynamic.[29] For each replacement, we randomly choose two agents, and the agent with the higher score is copied into the agent with the lower score.

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

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

When we raise group size to $n = 10$, however, the picture is quite different. The first graph in Figure 1 illustrates the case with an error rate of $\epsilon = 5$ percent. Note that even with this relatively small group size, the level of cooperation falls to very low levels. Lowering the error rate to $\epsilon = 0.5$ percent, as in the second graph in Figure 1, we see that the level of cooperation becomes high, but the efficiency of cooperation is only about 17 percent. This is because cooperation is signaled as defection between some pairs of agents with a probability close to 40 percent. Only when we set the error level to $\epsilon = 0.1$ percent, as in Figure 1, do we achieve a high level of efficiency, the probability of an agent receiving a defection signal when in fact all are cooperating now falling below 10 percent. Since this low error level also allows a high level of tolerance, defections become quite rare. However, a 0.1 percent error rate is implausibly low.

3. The concept of rational behavior

Can the behavior of 'real people' be modeled as though they were 'maximizing some utility function'? Despite Binmore's denial, the answer is almost surely in the affirmative. It is often thought that the idea of maximizing applies only when extremely stringent rationality and complete information conditions are satisfied. However, the model can be shown to apply over any domain in which (1) the agent has *transitive preferences* (in the sense that if an agent prefers A to B and prefers B to C, then the agent prefers A to C) and (2) the agent can *trade off among outcomes* (in the sense that for any finite set of payoffs $A_1, ..., A_n$, if A_1 is the least preferred and A_n the most preferred, then for any A_i, $1 \le i \le n$, there is a probability p_i, $0 \le p_i \le 1$ such that the agent is indifferent between A_i and a lottery that pays A_1 with a probability p_i and pays A_n with a probability $1 - p_i$).[30]

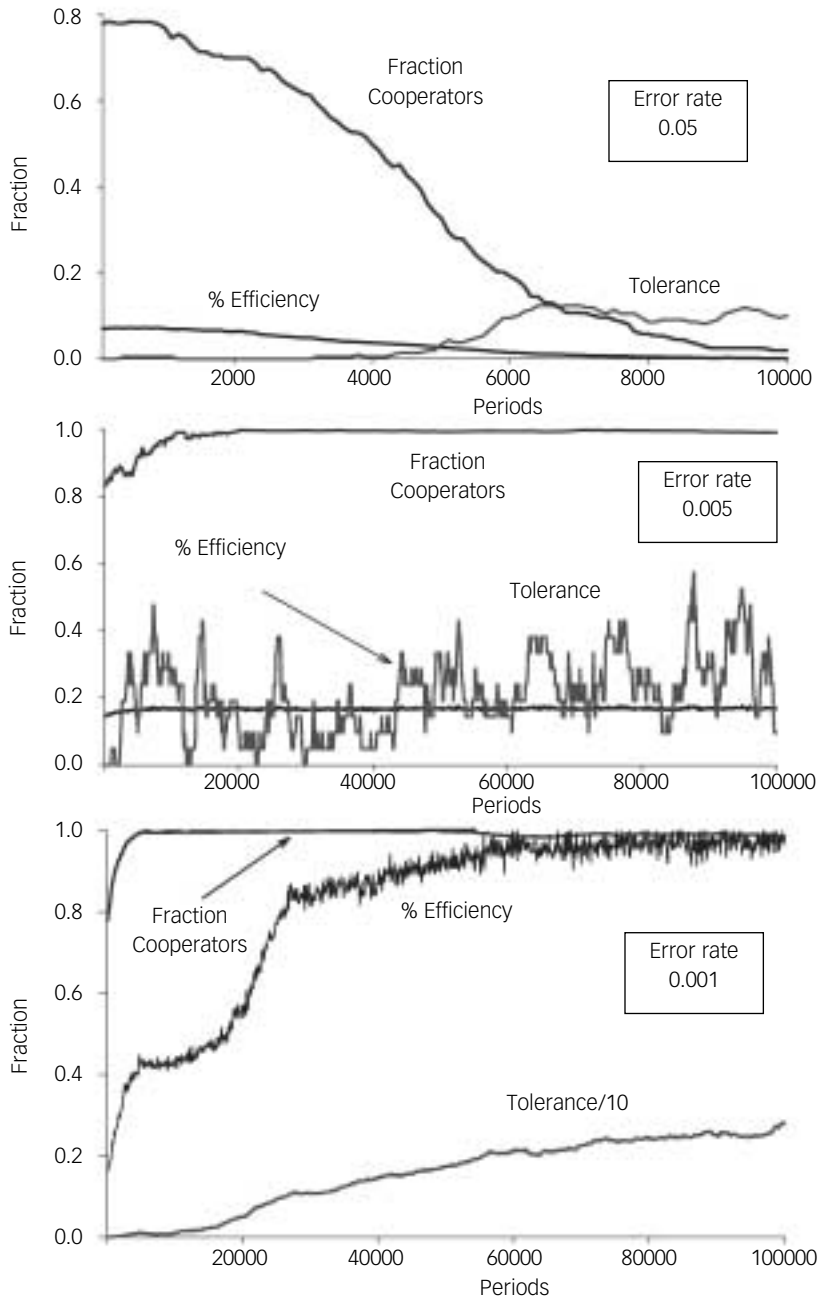


Figure 1 **Simulation of the Bhaskar-Obara model with group size *n* = 10, with model parameters as described in the text**

These assumptions are hardly demanding. When applicable, assuming maximization subject to constraints strongly enhances explanatory power, even in areas that have traditionally abjured maximization.[31]

Transitivity of preferences is ubiquitous because *any evolved life form is likely to conform to these conditions*. This is because biological agents possess an evolved, genetically rooted set of routines, involving needs, drives, pleasures, and pains that determine how to respond to internal events (for example, hunger) and external circumstances (for example, temperature) so as to promote their long-term fitness. Since each combination of internal and external states will be associated with a real number representing a fitness value, and since the real numbers form an ordered field, a biological agent's preference function will tend to be transitive. Evolutionary forces ensure that, under constant environmental conditions, maximizing this preference function will in fact come close to maximizing the agent's fitness. But even when environmental conditions so change that an agent's preferences no longer conform to fitness maximization, the preferences themselves will remain transitive, so that the agent's behavior can be modeled as utility maximization.

The notion that agents maximize utility does not require that agents be self-regarding, since there is no connection between the transitivity of preferences and the content of preferences. Indeed, one can apply standard choice theory, including the derivation of demand curves, plotting concave indifference curves, and finding price elasticities for such preferences as charitable giving and punitive retribution.[32] There is thus nothing 'irrational' about such behavior.

As Binmore suggests, broadening the model of the individual maximizing a utility function beyond its traditional form in neoclassical economics runs the risk of developing unverifiable and *post hoc* theories, as our ability to theorize outpaces our ability to test theories. To avoid this, we must expand the use of controlled experiments and field data. Often we find that the appropriate experimental design can generate new data to distinguish among models that are equally powerful in explaining the existing data.[33] It is to this issue that I now turn.

4. Other-regarding behavior in humans

In this section, I will describe an elegant experiment carried out by Ernst Fehr, Simon Gächter, and Georg Kirchsteiger in 1997.[34] This is one of a host of experiments exhibiting what I have termed 'strong reciprocity'.[35] Strong reciprocity is a predisposition to cooperate with others, and to punish those who violate the norms of cooperation, at personal cost, even when it is implausible to expect that these costs will be repaid.[36] The experimenters divided a group of 141 subjects into a set of 'employers' and a larger set of 'employees' (the experimenters used socially neutral terms). The rules of the game may be stated as follows. If an employer hires an employee who provides effort e and receives a wage w, the

employer's payoff π; is 100 times effort e minus wage w that he must pay the employee ($\pi = 100e - w$), where the wage is between zero and 100 ($0 \leq w \leq 100$) and effort between 0.1 and 1 ($0.1 \leq e \leq 1$). The payoff u to the employee is then the wage he receives minus a 'cost of effort', $c(e)$ ($u = w - c(e)$). The cost of effort schedule $c(e)$ is constructed by the experimenters such that supplying effort e = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 costs the employee $c(e)$ = 0, 1, 2, 4, 6, 8, 10, 12, 15, and 18, respectively. All payoffs are converted into real money that the subjects are paid at the end of the experimental session.

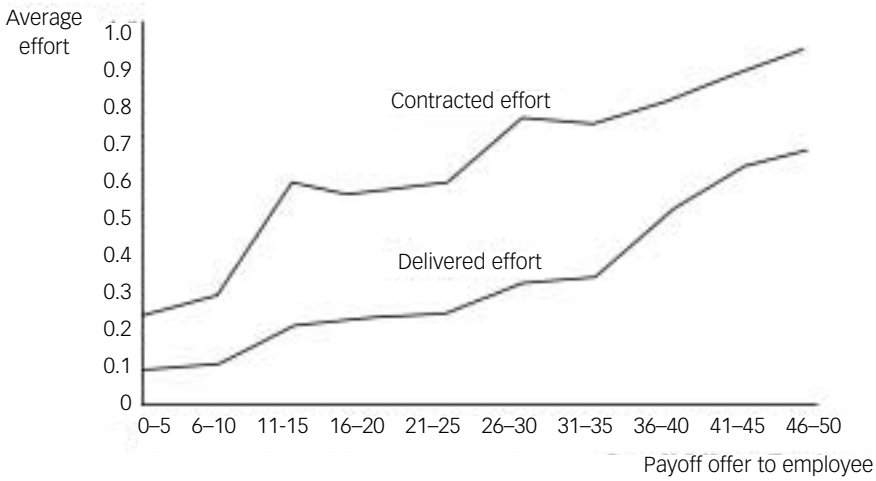
The sequence of actions is as follows. The employer first offers a 'contract' specifying a wage w and a desired amount of effort e^*. A contract is made with the first employee who agrees to these terms. An employer can make a contract (w, e^*) with at most one employee. The employee who agrees to these terms receives the wage w and supplies an effort level e, which *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 (unrepeated) 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 integral wage offers are permitted). The employee will accept this offer, and will set e = 0.1. Since $c(0.1) = 0$, 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 strong reciprocity predispositions among 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 2. Similar results have been observed in Fehr, Kirchsteiger, and Riedl.[37]

Figure 2 also shows that, though most employees are strong reciprocators, at any wage rate there is still a significant gap between the amount of effort agreed upon and the amount actually delivered. This is not because there are a few 'bad apples' among the set of employees, but because only 26 percent of employees delivered the level of effort they promised. We conclude that strong reciprocators are inclined to compromise their morality to some extent, just as we might expect from daily experience.

The above evidence is compatible with the notion that the employers are purely self-regarding, since their beneficent behavior vis-a-vis their employees was effective in increasing employer profits. To see if employers are also strong reciprocators, following this round of experiments, the authors extended the game by allowing the employers to respond reciprocally to the *actual effort*



Source: Fehr, Gächter and Kirchsteiger (1997), 'Reciprocity as a Contract Enforcement Device: Experimental Evidence'.

Figure 2 **Relation of contracted and delivered effort to worker payoff (141 subjects)**

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, since they would not interact with the same worker a second time. However, 68 percent of the time employers punished employees that did not fulfill their contracts and 70 percent of the time employers rewarded employees who overfulfilled their contracts. Indeed, employers rewarded 41 percent 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 83 percent to 26 percent of the exchanges, and overfulfilled contracts rose from 3 percent to 38 percent of the total. Finally, allowing employers to reward and punish led to a 40 percent increase in net payoffs to all subjects, even when the payoff reductions resulting from employer punishment of employees were taken into account. Several researchers have predicted this general behavior on the basis of general, real-life social observation and field studies.[38] The laboratory results show that this behavior has a motivational basis in strong reciprocity and not simply long-term material self-interest.

We conclude from this study that the subjects who assume the role of 'employee' conform to internalized standards of reciprocity, even when they know 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' draw upon the internalized norm of rewarding good and punishing bad behavior when they are permitted to punish, and 'employees' expect this behavior and adjust their own effort levels accordingly.

I have developed the above experiment in detail to give the unfamiliar reader a flavor of experimental techniques. We may summarize other experimental results as follows. These results generally support the strong reciprocity model, and there is enough consistency in human behavior to support the notion that agents maximize a utility function that includes a taste for cooperating as well as a taste for punishing noncooperators.

In one-shot prisoner's dilemma experiments, the rate of cooperation is commonly between 40 percent and 60 percent.[39] Many subjects prefer the mutual cooperation outcome over the higher material payoff they would get by defecting on a cooperator. Moreover, many defect not because they are self-regarding, but to avoid risking being exploited by a selfish partner. For instance, Kiyonari, Tanida, and Yamagishi show that if subjects are informed that their partner has already moved and has cooperated, they are much more likely to cooperate than in the absence of such information.[40] In another one-shot game in which the subjects are mutually anonymous, the so-called ultimatum game,[41] one subject, called the 'proposer', is given a sum of money, say $10, and is instructed to offer any number of dollars, from $1 to $10, to a second subject, called the 'responder'. The responder can either accept the offer or reject it. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

Since the game is one shot and anonymous, a self-regarding responder will accept *any positive amount of money*. Knowing this, a self-regarding proposer will offer the minimum possible amount, say $1, and this will be accepted. However, when actually played, *the self-regarding outcome is almost never attained and rarely is approximated*. As many replications of this experiment in more than 30 countries have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts, 50 percent of the total generally being the modal offer. Respondents frequently reject offers of less than 30 percent.[42] The fact that positive offers are commonly rejected shows that respondents have other-regarding preferences, and the fact that most proposers offer $4 or $5 shows that either proposers have other-regarding preferences or they at least believe respondents have other-regarding preferences.

Are there other plausible interpretations of this behavior? One might suggest that subjects simply did not understand the game. This is not very plausible, because the game is extremely simple and experimenters generally require subjects to exhibit understanding before permitting them to participate. Moreover, if failure to understand were the problem, subjects who play several ultimatum

games in succession with different partners should eventually learn to accept any positive offer. In fact, generally, the rejection rate does not decline with repetition.

Other data support the notion that responders reject positive offers in order to punish an unfair proposer, and not because they are confused. For instance, in a variant of the game in which a responder rejection leads to the responder getting nothing, but allowing the proposer to keep the share he suggested for himself, respondents rarely reject offers, and proposers make considerably smaller (but still positive) offers. Also, consider the ultimatum game with one small change: the offer is not chosen by the proposer, but is generated by a computer, and the responder is told this fact. In this new situation, the rejection rate becomes very low, however small a share is offered to the responder.[43]

Another possibility is that other-regarding preferences emerge when the stakes are low, but would disappear with higher-stake games. Several researchers have tested this proposition, and found it not to hold.[44] Slonim and Roth, however, show that if the ultimatum game is repeated 10 times with different partners each time, there is a small, but significant tendency for rejections to decline when the stakes are very high (about 10 days' wages in Slovakia), but not otherwise.[45]

A third possibility (the one favored by Binmore) is that subjects are not used to one-shot, anonymous games, so they respond emotionally as they would in a repeated, non-anonymous interaction in everyday life. 'But why should responders get angry [when they are offered a small share of the pie]?' asks Binmore. 'I think they get angry', he says, 'because this is their habituated response to an unfair offer in the situations in which we encounter *ultimata* in real life. It is then almost never true that the game is one-shot.'[46]

This response is problematic in two ways. First, it is simply untrue that we are not used to one-shot, anonymous situations. Civility to others when riding on trains and planes, giving to charity, voting, attending public functions, and not to mention participating in spontaneous collective actions in support of a moral ideal are common and ubiquitous. Were people to act in a self-regarding manner in such situations, modern society could not function.

Second, even if Binmore's argument that we confuse one-shot with repeated games were correct, it would not lessen the importance of other-regarding preferences, given the importance of anonymous, unrepeated interactions in public life. But his argument is not correct. In fact, humans are well capable of distinguishing individuals with whom they are likely to have many future interactions from others with whom future interactions are less likely. Indeed, human experimental subjects cooperate much more if they expect frequent future interactions than if future interactions are likely to be infrequent.[47]

Binmore asserts that if the ultimatum game is repeated many times with the same subjects, responders eventually begin to accept lower offers. There are two problems with this argument. First, this behavior, if it exists, is very weak. The only study supporting it at all (Binmore gives no references) is Slonim and Roth,

discussed above.[48] Second, the desirability of a valued good or service can be expected to decline, according to traditional consumer theory. This is called 'satiation' or 'diminishing marginal utility'. The astonishing fact is that as far as we can tell from experiments, there is only a very weak tendency for satiation in the demand for being treated fairly.

According to what I have learned from experimental games, strong reciprocity is not a 'habituated response' that one can learn to abandon when it is materially welfare enhancing to do so. Binmore asserts that 'what the responder's bodies have to learn is that there isn't any point in getting angry with a stranger you are never going to meet again . . . the harder the knocks, the quicker we are conditioned with emotional responses that fit the actual game we are playing'.[49] Despots would, of course, love it were this in fact the human norm, for then simple material incentives, including the whip and the dungeon, would be completely adequate to keep subjects in line. But it is not true. Individuals who have rewarded those who help them, or hurt those who have hurt them, do not later regret having fallen prey to irrational emotionality. Rather, they generally affirm the morality of their behavior. The flaw in Binmore's argument is that there is no plausible mechanism that leads individuals to transform their emotional responses when they are in conflict with their material welfare. Even in the long run, when we might expect evolutionary forces to drive out emotions that conflict with self-interest, other evolutionary forces in humans conspire to maintain prosocial emotions intact.[50]

5. Programmable preferences and altruism

The distinctive contribution of sociological theory to understanding human behavior is *socialization theory*. A key tenet of socialization theory is that adults use their dominant cultural institutions to *program* society's values into the psyche of impressionable youth through the *internalization of norms*.[51] In the language of rational choice theory, internalized norms are accepted not as instruments toward and constraints upon achieving other ends, but rather as *arguments in the preference function that the individual maximizes*. A variety of prosocial emotions then come into play, including prominently *shame*, *guilt*, and *empathy*, directly biasing individual choices in prosocial directions.[52]

By noting that individuals internalize norms, we can derive altruistic behavior without requiring intergroup competition of the sort needed by group selection models of altruistic behavior. This is desirable because biologists have shown that genetic group selection is very difficult to sustain unless genetic relatedness among group members is very high.[53] In the biological literature, *group selection* applied to a trait has tended to mean that the trait suffers a within-group fitness deficit, but nevertheless grows in the population because groups in which the trait is prevalent outcompete other groups in which the trait has a low frequency.[54] There is a weaker sense of group selection, in which a trait is associated with, or

facilitates, a certain group structure and groups with this structure outcompete groups without the structure, but the selected trait does not have a within-group fitness disadvantage. The biological critiques of group selection do not apply to this weaker notion. The internalization of norms involves group selection only in this weaker, uncontroversial sense.

Why do we have the generalized capacity to internalize norms? This capacity is certainly unusual in the world of living beings – something akin to the capacity of a digital computer to be programmed. From a biological standpoint, internalization may be an elaboration of imprinting and imitation mechanisms found in several species of birds and mammals, but its highly developed form in humans indicates it probably had great adaptive value during our evolutionary emergence as a species. Moreover, from an economic standpoint, the everyday observation that people who exhibit a strongly internalized moral code lead happier and more fulfilled lives than those who subject all actions to a narrow calculation of personal costs and benefits of norm compliance suggests it might not be 'rational' to be self-regarding.

Gintis shows that *if* internalization of *some* norms is personally fitness enhancing (for example, preparing for the future and having good personal hygiene, positive work habits, and control of one's emotions), *then* genes promoting the capacity to internalize can evolve. Given this genetic capacity, as we have seen above, altruistic norms will be internalized as well, provided their fitness costs are not excessive. In effect, altruism 'hitchhikes' on the personal fitness-enhancing capacity of norm internalization.[55]

The internalization of norms is adaptive because it facilitates the transformation of individual drives, needs, desires, and pleasures (arguments in the human preference function) into forms that are closely aligned with fitness maximization in a highly variable cultural environment. We humans have a 'primordial' preference function that does not well serve our fitness interests, and which is more or less successfully 'overridden' by our internalized norms. This primordial preference function knows little of 'thinking ahead', but rather satisfies immediate desires. Lying, cheating, killing, stealing, and satisfying short-term bodily needs (wrath, lust, greed, gluttony, and sloth) are all actions that produce immediate pleasure at the expense of our long-run well-being.[56]

This evolutionary argument is meant to apply to the long period in the Pleistocene during which the human character was formed. Social change since the agricultural revolution some 10,000 years ago has been far too swift to permit even the internalization of norms to produce a close fit between preferences and fitness. Indeed, with the advent of modern societies, the internalization of norms has been systematically diverted from *fitness* (expected number of offspring) to *welfare* (net degree of contentment) maximization. This is precisely what we would expect when humans obtain control over the content of ethical norms. Indeed, this *misfit* between welfare and fitness is doubtless a necessary precondition for civilization and a high level of per capita income because, were

we fitness maximizers, every technical advance would have been accompanied by an equivalent increase in the rate of population growth, thus nullifying its contribution to human welfare, as predicted long ago by Thomas Malthus. The demographic transition, which has led to dramatically reduced human birth rates throughout most of the world, is a testimonial to the gap between welfare and fitness.

6. Charity

The issue of 'charity' is not dealt with explicitly in *Natural Justice*. Binmore does note in Chapter 11 that we can expect a person's social index to decline as the person's social status increases. If the poor are also low status, this implies that there will be a tendency to transfer resources from the rich to the poor. Thus, charity can be explained using Binmore's principles of natural justice. But, as I suggest below, charity can be much better explained by a model of other-regarding preferences, including strong reciprocity.

In the advanced economies, a substantial fraction of total income is regularly transferred from the better off to the less well off, and the governments that preside over these transfers are regularly endorsed by their publics.[57] The modern welfare state is thus the most significant example in human history of a voluntary egalitarian redistribution of income among total strangers. What accounts for its popular support?

I suggest below that a compelling case can be made that people support the welfare state because it conforms to deeply held norms of reciprocity and conditional obligations to others. Abundant evidence from across the social sciences shows that strong reciprocity governs charitable giving: when people blame the poor for their poverty, they support less redistribution than when they believe that the poor are poor through no fault of their own.[58] Concern about the 'undeserving poor' is pronounced in the USA, but is far from absent in Europe. Fong, Bowles, and Gintis show that in 12 European countries those who say that poverty is the result of laziness support less government redistribution and are less concerned about unemployment, poverty, and inequality than those who do not.[59]

Strong reciprocity is at the heart of taxpayer attitudes toward poverty relief. To see this, consider the 1998 Gallup Poll Social Audit Survey, 'Haves and Have-Nots: Perceptions of Fairness and Opportunity', analyzed in Fong, Bowles, and Gintis.[60] This study found that those who say that bad luck alone causes poverty are 0.50 standard deviation higher in their support for redistribution than those who think lack of effort alone causes poverty. Those who think that good luck alone causes wealth are 0.39 standard deviation higher on the support-for-redistribution scale than those who think effort alone causes wealth, and people who respond that there is plenty of opportunity in the USA to get ahead scored 0.42 standard deviation lower in their support for redistribution than people who do not think there is plenty of opportunity.

Even more convincing evidence on this point comes from an experiment including actual welfare recipients.[61] There were no disincentive costs at all in this experiment, yet, student subjects gave more to those welfare recipients with a stronger work commitment. These results lend strong support to previously made hypotheses about well-known patterns in survey data. Heclo reports that 81 percent of survey respondents favor public funding for childcare if the mother is a widow who is trying to support three children, while only 15 percent favor public funding when the mother has never married and is not interested in working.[62] Heclo also reports the results of a survey in which the wording of a question about support for public redistribution was manipulated so that some subjects were asked about spending on 'welfare', while others were asked about spending on 'assistance for the poor' or 'caring for the poor'. In that experiment, 41 percent of respondents stated that there is too much spending on welfare and 25 percent stated that there is too little. By contrast, only 11 percent and 7 percent of the respondents said that there is too much spending on assistance for and caring for the poor, respectively, and 64 percent and 69 percent said that there is too little spending on assistance for and caring for the poor, respectively. In a similar vein, Page and Shapiro report that support for social security spending has been very high and stable over time, while support for spending on welfare has been consistently low.[63] The interpretation commonly given for findings such as these is that people are less generous to recipients who they think are not working when they could and should be, or who are otherwise considered to be in questionable moral standing.[64] I have shown that these findings cannot be explained away by a fuller and more rigorous account of self-interest.

7. Conclusion

While I have great sympathy for Binmore's project of applying empirical data and economic theory to understanding justice in human society, I believe he has relied heavily in carrying out this project on an aspect of economic theory that has ~~had~~ little empirical relevance. As I have tried to show, repeated game theory has had many brilliant successes, but these do not include having developed a plausible model of prosocial behavior in moderate-sized groups of unrelated, self-regarding agents. Nor have its proponents ever submitted repeated game models to empirical testing, and hence their acceptance of these models has been based on criteria other than empirical relevance.

I also do not believe that the moral rules deployed in human societies are correctly derived by applying local cultural-social indices to the Nash bargaining problem. Rather, I think that human beings are constituted, by virtue of their evolutionary history, to behave as altruistic cooperators and punishers whose egalitarian predilections stem from a long history of enforced egalitarianism in the hunter-gatherer societies from which modern humanity emerged.

Appendix A: Imperfect public signals and trigger strategies

We assume that agents can shirk when intending to cooperate, but cannot cooperate when intending to shirk. Since shirking occurs with positive probability in each period, shirking will eventually occur with a probability of 1 (if the error rate is $\epsilon > 0$, then the expected time until shirking occurs is $1/\epsilon$). Since the punishment phase will certainly occur at some point, it now becomes cost-effective to revert to universal defection for the minimum number of periods (say, k) that just make it unprofitable to shirk purposely. The value of cooperating when all other members cooperate is now given by the recursion

$$v_c = b(1-\epsilon) - c + \delta(1-\epsilon)^n v_c + (1-(1-\epsilon)^n)\delta^{k+1} v_c\,, \tag{1A}$$

which gives

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

The present value of defecting is now $v_d = b(1-\epsilon) + v_c\,\delta^{k+1}$. By the one-shot deviation principle, cooperation is Nash sub-game perfect if and only if $v_c \geq v_d$, which simplifies to

$$\frac{b}{c}(1-\epsilon)^{n+1} \geq \frac{1-\delta^{k+1}}{\delta(1-\delta^k)}\,. \tag{3A}$$

It is easy to check that the right-hand side of Equation 3A is always greater than $1/\delta$, so for any b, c, and e, when n becomes large enough, the condition fails for any $\delta < 1$. That is, the cooperative equilibrium cannot be sustained, no matter how patient the group members. Thus, no matter how small the probability ϵ, if the group is sufficiently large, cooperation cannot be sustained. It is also easy to check that the total discounted payoff to members as $k \to \infty$ becomes

$$v_c = b(1-\epsilon) - c + \delta(1-\epsilon)^n\,, \tag{4A}$$

which simplifies to

$$v_c = \frac{b(1-\epsilon) - c}{1 - \delta(1-\epsilon)^n}\,, \tag{5A}$$

and which is close to the one-shot payoff to cooperation when n is large.

For example, suppose $n = 15$, $b/c = 1.5$, $\delta = 0.95$, and $\epsilon = 2.0$ percent. Then we must set $k = 19$ for a cooperative equilibrium, the punishment stage occurs with an 18 percent probability, and the present value of the game is 2.3; whereas, if cooperation could be costlessly enforced, the present value of the game would be 9.4. Also, if the error rate reaches about 2.2 percent, cooperation cannot be maintained at all.

Appendix B: Imperfect public signals and directed punishment

If all agents cooperate and punish, the rate of defection observed will be ϵn, so the mean number of punishment events per period per agent will be ϵn, and hence the mean number of punishment events per period will be ϵn^2. The mean number of signaled failures to punish will be $n^2\epsilon^2$, which will require $n^2\epsilon^2$ punishments of non-punishers. A fraction $n^3\epsilon^3$ of these will fail, on average, requiring additional rounds of punishment. All in all, the expected number of punishment events per period will be

$$n\epsilon + n^2\epsilon^2 + n^3\epsilon^3 + \ldots = \frac{n\epsilon}{1 - n\epsilon} \cdot \tag{1B}$$

Thus, as the error rate approaches $1/n$, the number of expected punishments becomes infinite. It is easy to check that for punishment to be effective, p must satisfy

$$p \geq \frac{(c_p - c)\epsilon + c/n}{1 - \epsilon(n + 1)}, \tag{2B}$$

and that using this value for p, v_c is positive only if, approximately,

$$n\epsilon < \frac{1}{\frac{b}{n} + \frac{c_p}{b-c}} \cdot \tag{3B}$$

We conclude that directed punishment can be effective for small groups, since with reasonable error rates, ϵn can be held to a low level. However, for groups of 10 or more members, directed punishment is not likely to be effective at reasonable error rates, especially since the assumption of public signaling is rarely plausible in such groups.

notes

I would like to thank the associate editor and an anonymous reviewer for helpful comments, and the John D. and Catherine T. MacArthur Foundation for research support.

1. Herbert Gintis, *Game Theory Evolving* (Princeton, NJ: Princeton University Press, 2000).
2. 'What can be said can be said clearly; what one cannot speak about, one should be silent about.'
3. Traditional philosophical ethicists, says Binmore, 'have no more access to some noumenal world of moral absolutes than the boy who delivers the newspapers'. See Ken Binmore, *Natural Justice* (Oxford: Oxford University Press, 2005), p. 1.
4. 'Kant never provides a genuine defense of his claims . . . his categorical imperative is simply a grandiloquent rendering of folk wisdom.' See Binmore, *Natural Justice*, p. ix.

5. Edward O. Wilson, *Sociobiology: The New Synthesis* (Cambridge, MA: Harvard University Press, 1975).
6. Throughout human history, up to about 10,000 years ago with the emergence of sedentary agriculture and private property, there was no institutional means by which a minority could impose rules on the majority, so, as Binmore notes, fairness probably flourished, as is the case in virtually every contemporary hunter-gatherer group.
7. See Binmore, *Natural Justice*, Section 4.6.
8. Ibid., p. 19.
9. Throughout this article, I will use the term 'self-regarding' rather than the more common term 'self-interested'. This is to avoid the confusion stemming from the fact that other-regarding agents may be self-interested. For instance, if I care about the poor, then contributing to charity for the poor may be self-interested, despite the fact that it is not self-regarding.
10. Binmore, *Natural Justice*, p. 75.
11. Ibid.
12. Herbert Gintis, 'Strong Reciprocity and Human Sociality', *Journal of Theoretical Biology* 206 (2000): 169–79; Herbert Gintis, Samuel Bowles, Robert Boyd and Ernest Fehr (eds) *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life* (Cambridge, MA: MIT Press, 2005).
13. Ken Binmore, *Game Theory and the Social Contract: Playing Fair* (Cambridge, MA: MIT Press, 1994); Ken Binmore, *Game Theory and the Social Contract: Just Playing* (Cambridge, MA: MIT Press, 1998).
14. Drew Fudenberg and Eric Maskin, 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information', *Econometrica* 54 (1986): 533–54; Drew Fudenberg, David K. Levine and Eric Maskin, 'The Folk Theorem with Imperfect Public Information', *Econometrica* 62 (1994): 997–1039.
15. Binmore may have ignored this problem because he generally treats cooperation in the context of two-person games, where the weaknesses of traditional repeated game theory are at least evident. The success of our species, however, lies in our ability to cooperate in moderately large groups that somehow manage to solve the free-rider problem.
16. By a 'signal' I mean the indication each group member has concerning the cooperate/defect behavior of each other member. A *public* signal is one shared by all group members, so if one member receives a signal that another defected, all other members received the same signal. A *private* signal is one which need not be shared by other members: I may observe another member shirking, but others in the group do not receive this signal.
17. Reinhard Selten, 'A Note on Evolutionarily Stable Strategies in Asymmetric Animal Conflicts', *Journal of Theoretical Biology* 84 (1980): 93–101.
18. Gintis, 'Strong Reciprocity and Human Sociality', pp. 169–79; Hillard Kaplan and Michael Gurven, 'The Natural History of Human Food Sharing and Cooperation: A Review and a New Multi-Individual Approach to the Negotiation of Norms', in *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life*, edited by Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr (Cambridge, MA: MIT Press, 2004); Arthur Robson and Hillard Kaplan, 'The Evolution of Human Life Expectancy and Intelligence in Hunter-Gatherer Economies', *American Economic Review* 93 (2003): 150–69.

19. David Laibson, 'Golden Eggs and Hyperbolic Discounting', *Quarterly Journal of Economics* 112 (1997): 443–77.
20. Gintis, *Game Theory Evolving*.
21. An exception is workers striking for a period of time to protest their employer's violation of an agreed-upon work rule. In this example, however, the inefficiency of trigger strategies is largely absent, since we have in effect a two-player game between the employer and the workers' union.
22. Fudenberg and Maskin, 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information', pp. 533–54.
23. It is frequently suggested that the case of 'insiders' punishing their members for fraternizing with 'outsiders' is a case of second-order punishment. But, in an insider/outsider situation, the outsiders are not being punished – they are simply members of other groups. If the norm of a group is to refrain from fraternizing with members of other groups, defectors from this norm may be punished. But this is clearly a case of first-order punishment. Second-order punishment, in this case, would be to punish those who refuse to punish those who fraternize with outsiders.
24. Fudenberg and Maskin, 'The Folk Theorem in Repeated Games with Discounting or with Incomplete Information', pp. 533–54.
25. Dilip Abreu, David Pearce and Ennio Stacchetti, 'Toward a Theory of Discounted Repeated Games with Imperfect Monitoring', *Econometrica* 58 (1990): 1041–63; Fudenberg, Levine and Maskin, 'The Folk Theorem with Imperfect Public Information', pp. 997–1039.
26. Robert Sugden, *The Economics of Rights, Co-operation and Welfare* (Oxford: Basil Blackwell, 1986); Robert Boyd, 'Mistakes Allow Evolutionary Stability in the Repeated Prisoner's Dilemma Game', *Journal of Theoretical Biology* 136 (1989): 47–56.
27. See Michihiro Kandori, 'Introduction to Repeated Games with Private Monitoring', *Journal of Economic Theory* 102 (2002): 1–15 for a technical overview. Important contributions to this research agenda include Tadashi Sekiguchi, 'Efficiency in Repeated Prisoner's Dilemma with Private Monitoring', *Journal of Economic Theory* 76 (1997): 345–61; Michele Piccione, 'The Repeated Prisoner's Dilemma with Imperfect Private Monitoring', *Journal of Economic Theory* 102 (2002): 70–83; Jeffrey C. Ely and Juuso Välimäki, 'A Robust Folk Theorem for the Prisoner's Dilemma', *Journal of Economic Theory* 102 (2002): 84–105; V. Bhaskar and Ichiro Obara, 'Communication in the Repeated Prisoner's Dilemma with Private Monitoring', *Journal of Economic Theory* 102 (2002): 40–69; Elchanan Ben-Porath and Michael Kahneman, 'Communication in Repeated Games with Private Monitoring', *Journal of Economic Theory* 70 (1996): 281–97.
28. Bhaskar and Obara, 'Communication in the Repeated Prisoner's Dilemma with Private Monitoring', pp. 40–69.
29. For details of the replicator dynamic, see P. Taylor and L. Jonker, 'Evolutionarily Stable Strategies and Game Dynamics', *Mathematical Biosciences* 40 (1978): 145–56; Gintis, *Game Theory Evolving*.
30. David M. Kreps, *A Course in Microeconomic Theory* (Princeton, NJ: Princeton University Press, 1990).
31. James S. Coleman, *Foundations of Social Theory* (Cambridge, MA: Belknap, 1990); Peter Kollock, 'Transforming Social Dilemmas: Group Identity and Cooperation', in

Modeling Rational and Moral Agents, edited by Peter Danielson (Oxford: Oxford University Press, 1997); Michael Hechter and Satoshi Kanazawa, 'Sociological Rational Choice', *Annual Review of Sociology* 23 (1997): 199–214.

32. James Andreoni and John H. Miller, 'Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism', *Econometrica* 70 (2002): 737–53.
33. Amos Tversky and Daniel Kahneman, 'Loss Aversion in Riskless Choice: A Reference-Dependent Model', *Quarterly Journal of Economics* 106 (1981): 1039–61; Toko Kiyonari, Shigehito Tanida and Toshio Yamagishi, 'Social Exchange and Reciprocity: Confusion or a Heuristic?', *Evolution and Human Behavior* 21 (2000): 411–27.
34. Ernst Fehr, Simon Gächter and Georg Kirchsteiger, 'Reciprocity as a Contract Enforcement Device: Experimental Evidence', *Econometrica* 65 (1997): 833–60.
35. Gintis, 'Strong Reciprocity and Human Sociality', pp. 169–79.
36. For other such experiments, see Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr, 'Explaining Altruistic Behavior in Humans', *Evolution and Human Behavior* 24 (2003): 153–72.
37. Ernst Fehr, Georg Kirchsteiger and Arno Riedl, 'Does Fairness Prevent Market Clearing?' *Quarterly Journal of Economics* 108 (1993): 437–59; Ernst Fehr, Georg Kirchsteiger and Arno Riedl, 'Gift Exchange and Reciprocity in Competitive Experimental Markets', *European Economic Review* 42 (1998): 1–34.
38. See George Homans, *Social Behavior: Its Elementary Forms* (New York: Harcourt Brace, 1961); Peter Blau, *Exchange and Power in Social Life* (New York: John Wiley, 1964); George A. Akerlof, 'Labor Contracts as Partial Gift Exchange', *Quarterly Journal of Economics* 97 (1982): 543–69.
39. Ernst Fehr and Urs Fischbacher, 'Why Social Preferences Matter', in *Nobel Symposium on Behavioral and Experimental Economics* (2001); Colin Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton, NJ: Princeton University Press, 2003).
40. Kiyonari, Tanida and Yamagishi, 'Social Exchange and Reciprocity: Confusion or a Heuristic?', pp. 411–27.
41. Richard Thaler, 'Anomalies: The Ultimatum Game', *Journal of Economic Perspectives* 2 (1988): 195–206.
42. Alvin E. Roth, 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 (1991): 1068–95; Colin Camerer and Richard Thaler, 'Ultimatums, Dictators, and Manners', *Journal of Economic Perspectives* 9 (1995): 209–19.
43. Sally Blount, 'When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences', *Organizational Behavior and Human Decision Processes* 63 (1995): 131–44.
44. Roth, Prasnikar, Okuno-Fujiwara and Zamir, 'Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study', pp. 1068–95; Elizabeth Hoffman, Kevin McCabe and Vernon L. Smith, 'On Expectations and the Monetary Stakes in Ultimatum Games', *International Journal of Game Theory* (1994): 289–302; Paul G. Straub and J. Keith Murnighan, 'An Experimental Investigation of the Ultimatum Game: Common Knowledge, Fairness, Expectations,

and Lowest Acceptable Offers', *Journal of Economic Behavior and Organization* 27 (1995): 345–64; Lisa A. Cameron, 'Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia', *Economic Inquiry* 37 (1999): 47–59.

45. Robert Slonim and Alvin E. Roth, 'Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic', *Econometrica* 66 (1998): 569–96.
46. Binmore, *Natural Justice*, p. 83.
47. Claudia Keser and Frans van Winden, 'Conditional Cooperation and Voluntary Contributions to Public Goods', *Scandinavian Journal of Economics* 102 (2000): 23–39.
48. Slonim and Roth, 'Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic', pp. 569–96.
49. Binmore, *Natural Justice*, p. 84.
50. Herbert Gintis, 'The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms', *Journal of Theoretical Biology* 220 (2003): 407–18; Samuel Bowles and Herbert Gintis, 'The Origins of Human Cooperation', in *The Genetic and Cultural Origins of Cooperation*, edited by Peter Hammerstein (Cambridge, MA: MIT Press, 2003).
51. Emile Durkheim, *Suicide: A Study in Sociology* (New York: Free Press, 1951); Joan E. Grusec and Leon Kuczynski, *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory* (New York: Wiley, 1997).
52. Bowles and Gintis, 'The Origins of Human Cooperation'.
53. William D. Hamilton, 'The Evolution of Altruistic Behavior', *American Naturalist* 96 (1963): 354–6; Scott A. Boorman and Paul Levitt, *The Genetics of Altruism* (New York: Academic Press, 1980); John Maynard Smith, 'Group Selection', *Quarterly Review of Biology* 51 (1976): 277–83.
54. David Sloan Wilson, *The Natural Selection of Populations and Communities* (Menlo Park, CA: Benjamin Cummings, 1980).
55. Gintis, 'The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms', pp. 407–18.
56. George F. Loewenstein, 'Out of Control: Visceral Influences on Behavior', *Organizational Behavior and Human Decision Processes* 65 (1996): 272–92; Laibson, 'Golden Eggs and Hyperbolic Discounting', pp. 443–77.
57. A.B. Atkinson, *The Economic Consequences of Rolling Back the Welfare State* (Cambridge, MA: MIT Press, 1999).
58. John B. Williamson, 'Beliefs about the Motivation of the Poor and Attitudes Toward Poverty Policy', *Social Problems* 21 (1974): 734–47; Hugh Heclo, 'The Political Foundations of Antipoverty Policy', in *Fighting Poverty: What Works and What Doesn't*, edited by Sheldon H. Danziger and Daniel H. Weinberg (Cambridge, MA: Harvard University Press, 1986), pp. 312–41; Steve Farkas and Jean Robinson, *The Values we Live By: What Americans Want from Welfare Reform* (New York: Public Agenda, 1996); Martin Gilens, *Why Americans Hate Welfare* (Chicago, IL: University of Chicago Press, 1999); David Miller, *Principles of Social Justice* (Cambridge, MA: Harvard University Press, 1999).
59. Christina M. Fong, Samuel Bowles and Herbert Gintis, 'Reciprocity and the Welfare State', in *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life*, edited by Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr (Cambridge, MA: MIT Press, 2003).

60. Ibid.
61. Ibid.
62. Heclo, 'The Political Foundations of Antipoverty Policy', pp. 312–41.
63. Benjamin Page and Robert Shapiro, *The Rational Public: Fifty Years of Trends in American's Policy Preferences* (Chicago, IL and London: University of Chicago Press, 1992).
64. Heclo, 'The Political Foundations of Antipoverty Policy', pp. 312–41; Gilens, *Why Americans Hate Welfare*; Page and Shapiro, *The Rational Public: Fifty Years of Trends in American's Policy Preferences*.