ppe politics, philosophy & economics article

© SAGE Publications Ltd
London
Thousand Oaks, CA
and New Delhi

1470-594X
200602 5(1) 81–96

Why do people cooperate?

Ken Binmore
University College, London, UK

abstract Can people be relied upon to be nice to each other? Thomas Hobbes famously did not think so, but his view that rational cooperation does not require that people be nice has never been popular. The debate has continued to simmer since Joseph Butler took up the Hobbist gauntlet in 1725. This article defends the modern version of Hobbism derived largely from game theory against a new school of Butlerians who call themselves behavioral economists. It is agreed that the experimental evidence supports the claim that most people will often make small sacrifices on behalf of others and that a few will sometimes make big sacrifices, but that the larger claims made by contemporary Butlerians lack genuine support.

keywords natural justice, fairness, social norms, game theory, behavioral economics

Introduction

I have a lot in common with the three scholars invited to comment on my *Natural Justice* in this issue.[1] All of us are skeptical about the existence of the various supernatural agencies whose authority is commonly invoked by those who think there are universal moral absolutes. Our skepticism extends not only to gods and demons, but to such modern substitutes as practical reason, moral intuition, or natural law. In seeking to understand how human morality works, we think it appropriate to look, instead, at what the theory of evolution has to tell us about the prehistory of social animals such as ourselves. We therefore read biology books. We study anthropological reports. We conduct laboratory experiments. In brief, we try to approach the subject of human morality scientifically.

We cannot hope to find favor with traditional moral pundits. Our moral relativism is regularly denounced by the current Pope. Academic moral philosophers feel that our lack of respect for the likes of Plato or Immanuel Kant excuses them from the necessity of listening to what we have to say. Postmodern scholars think

DOI: 10.1177/1470594X06060620

Ken Binmore is Professor of Economics, Emeritus, at University College London, Drayton House, Gordon Street, London, WC1H 0AN, UK [email: k.binmore@ucl.ac.uk]

our reliance on scientific method is hopelessly naive. Social scientists of the old school are dismayed by our rejection of the dogma that the human mind is a blank slate on which anything whatever can be written. No matter how much we insist on the importance of cultural evolution, they therefore seek to discredit our efforts by calling us genetic determinists or sociobiological fascists. Everybody else shares the general distrust of the supposedly mean-minded, money-grubbing misfits who choose to enter the economics profession.

What is our response to finding ourselves beset by this battery of entrenched opinion? Like all beleaguered academic minorities, we respond by playing our own little version of the prisoner's dilemma. Rather than confront the barbarians beneath our walls, we spend our time accusing each other of factual inaccuracies and obscure heresies. But who cares, for example, whether the reciprocity that we all agree is necessary to sustain cooperation in the human species is of the strong kind or the weak kind? Certainly not the public intellectuals whose endorsement of our joint enterprise would be so welcome!

But what can a game theorist do in the prisoner's dilemma (or any other game) but seek to make his best reply to the strategies chosen by the other players? It is in this spirit that I join Don Ross in defending the position I adopt in *Natural Justice* from the criticism of Herb Gintis and Paul Seabright. However, the fact that Ross and I think that our views will one day become the received doctrine of a faith that will eventually conquer the world should not be allowed to conceal the fact that all four of us are singing from essentially the same hymn sheet. In particular, if it should finally turn out that it is Gintis and Seabright who have the right end of the stick and Ross and I who are wrong, then substantial chunks of *Natural Justice* would need to be rewritten, but its essential message would still be the same.

Behavioralism

The discredited doctrines of B.F. Skinner are sufficiently far in the past that it is now possible to admit to being an economic behavioralist without being subjected to knee-jerk hostility. Behavioralists in my own camp run laboratory experiments to test the extent to which mainstream economic assumptions about human behavior work out in practice. Well-known practitioners include Charlie Plott, Al Roth, and the recent Nobel laureate, Vernon Smith. We find that game theory works fairly well most of the time, when the following three criteria are satisfied:

- the problem faced by the subjects is not too complex, and is presented to the subjects in a user-friendly style,
- the subjects are adequately incentivized, and
- the subjects have ample time for trial-and-error learning.

Under these conditions (and sometimes under less stringent conditions) Nash equilibrium predicts rather well most of the time, without departing from the naive assumption that players are seeking to maximize their average monetary gain.

A rival school of behavioral economists challenges this summary of the experimental situation. Well-known practitioners include Colin Camerer, Ernst Fehr, and the recent Nobel laureate, Danny Kahneman. In the case of the bargaining experiments with which I have mostly been concerned, they argue that the data can only be explained by denying that players are seeking to maximize average monetary gain. We must instead attribute their behavior to their holding 'other-regarding preferences'. It is to this school that Herb Gintis and Paul Seabright appeal in criticizing my *Natural Justice*.

The debate is by no means new. It reaches at least as far back as Joseph Butler's attempt in 1726 to rebut the claim of Thomas Hobbes that cooperation among human beings can largely be explained in terms of rational self-interest.[2] I have only been part of the debate for 20 years or so, but I must admit to a weary sense of déjà vu as new Butlerians continually revive arguments that old Hobbists like myself thought had finally been laid to rest.

Seabright's endorsement of claims made by the psychologists Owren and Bachorowski is a good example. The last time the idea that humans can reliably signal their strategic dispositions to others through involuntary facial expressions or otherwise was Frank's *Passions within Reason*.[3] Before that, the same idea was proposed in Gauthier's *Morals by Agreement*.[4] But the first on this particular scene was Charles Darwin, who considers the idea at length in his *Expression of the Emotions in Man and Animals*.[5] I will not repeat Darwin's reasons for rejecting the idea, which are much the same as those that game theorists give for rejecting what we call the Transparent Disposition Fallacy.[6] I will only observe that one simply has to open a newspaper to read innumerable accounts of successful deceit.

It seems to me that a similar appeal to any newspaper or history book ought also to be enough to dispose of the idea that we are all genetically programmed to make large sacrifices for the sake of our fellow men, but Butlerians prefer to ignore the historical record in favor of a carefully selected menu of laboratory experiments, and it seems necessary that I respond to their criticism in the same terms.

It should first be noted that I do not deny the existence of other-regarding preferences. On the contrary, I trace their origin in Chapter 7 of *Natural Justice* to Hamilton's rule, which expresses the extent to which animals should be expected to make sacrifices on behalf of their kin. Since everyone in an ancestral hunter-gatherer group would have been a relative of some kind, it is therefore not surprising that most of us seem to be wired up to be willing to make *small* sacrifices to help out total strangers. Some of us are occasionally willing to make *big* sacrifices, but only a tiny fraction of saints are willing to do so on a regular basis. These conclusions seem to me to be consistent both with the experimental results in dictator games and with our general experience of behavior in the real world.

However, Butlerians want to go further than this. Rather than seeing other-regarding behavior as a relatively small deviation from the maximization of monetary payoffs, they insist that the laboratory evidence supports the claim that the deviation is large. They make this claim seem plausible in three ways:

1. The immense amount of evidence that supports the mainstream position is glossed over. For example, Paul Seabright is unhappy with the emphasis I give to the vast amount of experimental data available on the prisoner's dilemma and other games that model the private provision of a public good. But there is nothing pathological about the prisoner's dilemma. Neither Gintis nor Seabright would presumably deny that the laboratory norm is that most *experienced* players end up playing close to a Nash equilibrium in monetary payoffs in most games.
2. Attention is concentrated on cases that are anomalous when compared with the overall literature. The reasons that Hobbists give for why these cases should be expected to be anomalous are overlooked. I give three examples that include the ultimatum game in Section 5.
3. The fact that mainstream economists who pay attention to experimental results emphasize the importance of trial-and-error adjustment is ignored. Results in which the subjects have no opportunity for learning at all are then quoted as though they undermine the mainstream position.

It is the third point to which I attach most importance. The evolutionary success of Homo sapiens is commonly attributed to our being a flexible species capable of adapting our behavior to new situations through individual learning and cultural evolution. To adopt a model of human behavior that minimizes our capacity for adapting to new situations is to fly in the face of this insight.

It is particularly dangerous if such a model is used in attempts to reform our institutions. Mainstream economists who use the theory of mechanism design follow the lead of Thomas Hobbes and David Hume by assuming that rogues or knaves will eventually appear to exploit any loopholes in a newly reformed institution. They therefore appeal to the Nash equilibrium concept when predicting how people will behave after they have seen some rogues and knaves in action. In so doing, they accord with conventional wisdom. Here, for example, is an IRS Commissioner, Mark Everett, explaining why a survey shows that the percentage of the public who thought it OK to cheat on their taxes was up from 11 percent to 17 percent over the previous five years: 'It's a basic sense of fairness. Somebody out there is complying with the law, and they see others doing things, and over time, they feel like chumps.'[7]

So the IRS continues to audit on the assumption that nearly everybody will learn to cheat if they do not provide adequate disincentives. But what if the IRS were to be persuaded that mainstream economists give undue emphasis to learning in games that model the private provision of public goods, and so ceased to audit tax returns? I think that everybody knows that there soon would not be enough money to fund the public goods and services on which we all depend.

Interpreting laboratory evidence

Nobody predicts that inexperienced or unmotivated laboratory subjects will play one of the Nash equilibria of a game. For example, in the prisoner's dilemma, somewhat more than 50 percent of all subjects cooperate when they play for the first time. How do Hobbists explain this behavior?

We think the most likely explanation is that the framing of the game triggers a social norm that the players are accustomed to using when going about their everyday affairs. We see such a social norm as an equilibrium selection device that has evolved to allow us to coordinate our behavior in one or more of the games we play in real life. To survive, it must select a Nash equilibrium of the real-life games for which it is adapted, but the strategy profile it selects need not be a Nash equilibrium of a game the subjects face in the laboratory. In the case of laboratory games such as the prisoner's dilemma that model the private provision of public goods, the relevant real-life game is an *indefinitely repeated* game, for which the folk theorem tells us that cooperation can be sustained as a Nash equilibrium by strategies that punish anyone who defects.

The book *Foundations of Human Sociality* describes an enterprise in which a bunch of anthropologists tried out some of the games popular with Butlerians in various traditional societies all round the world.[8] The editors (who include Herb Gintis) argue that the results constitute one more nail in the coffin of Hobbism, but the actual reports of the anthropologists tell another story. Here is Jean Ensminger commenting on why the Orma of Uganda contributed generously in her public goods game:

> When this game was first described to my research assistants, they immediately identified it as the '*harambee*' game, a Swahili word for the institution of village-level contributions for public goods projects such as building a school . . . I suggest that the Orma were more willing to trust their fellow villagers not to free ride in the Public Goods Game because they associated it with a learned and predictable institution. While the game had no punishment for free-riding associated with it, the analogous institution with which they are familiar does. A social norm had been established over the years with strict enforcement that mandates what to do in an exactly analogous situation. It is possible that this institution 'cued' a particular behavior in this game.[9]

The enforcement here is enforcement by the players themselves as envisaged in the folk theorem, and not external enforcement by the government. (Ensminger tells us that national or cross-regional attempts at *harambee* collections are corrupt in the manner that a Hobbist would predict.)

In a later article, Henrich et al. seem to accept Ensminger's interpretation when they write: 'Experimental play often reflects patterns of interaction found in everyday life.'[10] If so, it is a huge mistake to try to explain the behavior of the Orma in the public goods game on the hypothesis that their behavior is adapted to the game they played in Ensminger's makeshift laboratory. In particular, inventing other-regarding utility functions whose maximization leads to generous contribution in the public goods game is futile. Ensminger is suggesting that her subjects' behavior is adapted to the public goods game embedded in the *repeated* game that they play in real life, for which the folk theorem provides an explanation that does not require us to invent anything at all.

Of course, if subjects play a laboratory game repeatedly (against a new opponent each time), then Hobbists predict that their behavior will eventually diverge from the equilibrium of the repeated game they customarily play in real life to some equilibrium of the one-shot game they are actually playing in the laboratory.[11] The evidence that this is what actually happens in western societies is overwhelming. The huge number of experiments on the private provision of public goods was surveyed independently by Ledyard and Sally.[12] Both observed that 90 percent of subjects end up contributing nothing – a conclusion endorsed in Camerer's recent *Behavioral Game Theory*.[13]

To what extent does such trial-and-error learning occur in the societies studied by Henrich et al. in the *Foundations of Human Sociality*? I do not know the answer because no data on this subject is reported. My guess is that in most (perhaps all) of the experiments the subjects never played the same game twice, as is clearly the case in Ensminger's account of her experiment.

Revealed preference

Don Ross seeks to reconcile my views with those of Gintis and Seabright by reminding us that the theory of revealed preference tells us that any consistent behavior can be described by saying that the decision maker is behaving *as though* maximizing a utility function. Since *any* behavior can be made consistent by including enough parameters in the model, there is no problem in understanding how Butlerians are able to fit utility functions to the behavior of subjects whom I think have not yet learned to adapt their behavior to the laboratory game with which they are faced. But why should we not dismiss this activity as a tautological exercise?

I think the answer depends on whether the utility function Butlerians fit to their data allows them to predict the behavior of subjects in at least some new situations. Here the difference between Hobbists and Butlerians is sharpest. I do not think the Butlerians can claim any genuine success in this arena at all.

We know that, in most games, subjects' behavior changes over time as they adapt to a new game. In such games, we therefore cannot use a utility function fitted to the behavior of inexperienced subjects in order to predict their future behavior even in the same game. *Foundations of Human Sociality* by Henrich et al. shows that we cannot use a utility function fitted to the behavior of inexperienced subjects in one society to predict the behavior of inexperienced subjects in the same game in another society. For example, Michael Alvard tells us that:

> As the results in this volume show, people do not universally play fair. The question is no longer why people seem to have a preference for fairness. The question is now: do people behave more or less fairly in adaptive ways?[14]

Some Butlerians claim that we are nevertheless able to use a utility function fitted to the behavior of inexperienced subjects in one game to predict the behavior of inexperienced subjects from the same society in other games, but I know of no cases where such attempts at prediction are successful.

The theory of inequity aversion proposed by Fehr and Schmidt[15] is usually quoted in denial of this skeptical assessment. Fehr and Schmidt claim to have used data from ultimatum games to calibrate the parameters in the other-regarding utility function of their theory, and then employed the calibrated utility function to predict the data from experiments on other games. However, Shaked[16] has pointed out that this claim cannot possibly be true, because the data supposedly used to calibrate the parameters only restricts their range. When Fehr and Schmidt pick particular values of the parameters from within this range, they are therefore making use of information that they should have denied themselves.[17]

In fact, my own experimental work shows that no other-regarding preferences whatever can be made to fit the data in two-stage ultimatum games, unless they take account of more than the subjects' own monetary payoffs and those of their opponents.[18] It therefore does not matter what values one assigns to the parameters in Fehr and Schmidt's other-regarding utility function on the basis of data from one-stage ultimatum games, because the resulting utility function will not be able to predict the data even in two-stage ultimatum games.

In brief, I think the Butlerian attempt to use economic theory to explain the behavior of inexperienced subjects is no less a failure than the attempts of the old-style Chicago school that they delight in criticizing. It is not just neoclassical economics that fails in this endeavor, but the retro-classical economics that the Butlerians espouse.[19] My own view is that we waste our time trying to work out what utility function inexperienced subjects are maximizing, because it is not useful to model them as maximizing anything at all. Economics is not the answer to everything, because we do not automatically behave rationally when confronted with novel problems. Insofar as we ever behave rationally, it is largely because we have the capacity to learn.

No learning?

Butlerians sometimes justify their discounting of the importance that Hobbists attach to learning or adaptation by observing that they see little evidence of any learning or adaptation in their laboratory studies. I think the evidence does indeed support this claim, but there is a good reason why we see little evidence of learning in the anomalous games to which Butlerians confine their attention.

In Chapter 9 of *Natural Justice*, I follow Peter Singer in speaking of an *expanding circle* when describing the process by which an equilibrium selection device may come to be used to select an equilibrium in a class of games for which it did not originally evolve. It is this same process that I think explains why subjects do not always learn to abandon the social norm that is originally triggered in a laboratory game. The reason is simple. If the social norm happens to nominate behavior in the laboratory game that is close to a Nash equilibrium of the laboratory game, then the later convergence on a Nash equilibrium that Hobbists predict becomes redundant.

Butlerians evade this obvious criticism of their favored interpretation of the data by failing to draw attention to the full set of Nash equilibria of the games they study. In spite of our repeated denials, they insist, for example, that Hobbists have no choice but to predict convergence only on *subgame-perfect* equilibria. However, it is at least 20 years since economic theorists first generated examples which show that evolutionary processes can very easily converge on Nash equilibria that are not subgame perfect. Butlerians also fail to appreciate how often slight perturbations in the payoff structure of the games they study can create new Nash equilibria whose play generates behavior close to the observed data. I have chosen three examples that are commonly cited by Butlerians against the Hobbist position to illustrate both points.

Ultimatum game

After the prisoner's dilemma, the ultimatum game must be the example on which most experimental work has been done. I have myself been a part of the enormous industry that has devoted itself to confirming over and over again that subjects do not play the subgame-perfect equilibrium if they are modeled as maximizing their monetary payoffs. Nor does their behavior seem to change much during the 10 or so trials that are sufficient to get near a Nash equilibrium in the prisoner's dilemma.

Binmore et al.[20] consider evolutionary processes in the ultimatum game at some length, confirming that they can easily converge on something other than the subgame-perfect equilibrium of the game. The same article also confirms the obvious point that an evolutionary process is bound to operate very slowly in the neighborhood of a Nash equilibrium (because the evolutionary pressures at a Nash equilibrium are zero by definition). It then only remains to note a point that is seldom mentioned by Butlerians: that any split whatever of the money available for division in the ultimatum game corresponds to a Nash equilibrium.

We therefore do not have to look very far for a reason why the social norm that subjects bring into the laboratory is not eroded as the subjects gain experience. Any social norm whatever will nominate a split of the money that corresponds to a Nash equilibrium.[21]

Public goods game with punishment

A much-cited experiment of Fehr and Gächter[22] begins by confirming the standard results on games that model the private provision of public goods – after 10 periods of experience nearly all subjects 'free ride' by contributing nothing. They then allow the players to punish free riders after each game. To punish, a subject must pay a small cost that results in the chosen deviant being subjected to a much larger cost. In the version with punishment, subjects learn to cooperate after 10 periods or so, although it can never be part of a subgame-perfect equilibrium of the game with monetary payoffs to pay a cost to punish someone you will never encounter again.[23]

But even if one insists on looking only at subgame-perfect equilibria of the one-shot game, it is unnecessary to postulate more than a small other-regarding component in the subjects' utility functions to create a game with a cooperative equilibrium. For example, Jakub Steiner[24] offers a simple model in which the subjects feel just a little angry with free riders. He then describes an equilibrium in which only the worst free rider would get punished. The small cost of punishing then becomes tiny, because it is shared among all the punishers. But the punishment is enough to support an equilibrium without free riding in the one-shot game, because a player who is the only free rider will necessarily be the most guilty.

Gift exchange

Another much-cited experiment of Fehr et al.[25] can be thought of as modeling a competitive labor market in which the workers have the opportunity to reward employers who pay above the competitive rate by putting in more effort, even though the employer has no comeback if the worker just pockets the extra money and shirks. The finding is that workers do indeed reward generous employers with more effort – that they metaphorically 'exchange gifts'. In summarizing their data, Fehr et al. say:

> These results indicate that reciprocity motives may indeed be capable of driving a competitive experimental market permanently away from the competitive outcome.[26]

However, in making this claim, the authors fail to take account of the final-round effects evident in the data reported in the appendix to their article. In 16 of the 26 final rounds in which the worker has the opportunity to reciprocate, he does not. On the contrary, his effort is as small as it is possible for it to be.[27]

My own guess is that an understanding of what is really going on in the Fehr et al. experiment requires appealing to the contagion mechanism described by Kandori[28] for sustaining cooperative equilibria in infinitely repeated games played by small groups of anonymous agents. It is true that the game of Fehr et al. is only repeated a finite number of times, but a number of authors, including Reinhard Selten,[29] have shown that the folk theorem often still works in the laboratory when the number of repetitions is finite. The fact that cooperation tends to break down in the final rounds of these experiments adds some support to my conjecture, once it is revealed that the same holds true in the experiment of Fehr et al. I think this point is sufficiently important that it is worth describing a simple model that exhibits the contagion phenomenon without the need to postulate that more than a small proportion of subjects behave as though they actively like reciprocating.[30]

A contagion model

In this simplified model of an employment market, there are m employers and n workers, where $m < n$. Each of N periods begins with each employer independently publishing for all to see either a *high* wage or a *low* wage. The workers get a negative payoff from being unemployed, and so they compete to get employed. Each worker has an equal chance, and so the probability that any single worker finds employment in any given period is m/n. The matchings are entirely anonymous, so that long-term relationships between an employer and a worker are impossible.

For simplicity, we assume that a worker on a *low* wage automatically shirks. But a worker on a *high* wage can choose *high* or *low* effort. Both members of a matched pair receive a payoff of s if the wage is *low* (and so the worker shirks). Both receive b if the wage is *high* and the worker puts in *high* effort. The worker receives a payoff of 1 and the employer a payoff of 0 if the wage is *high* and the worker puts in *low* effort.

If $0 < s < b < 1$, then all Nash equilibria of the finitely repeated game require that the employer offers a *low* wage along the equilibrium path. But if an experiment were run using this game, presumably the subjects would learn to behave as in the more complicated market of Fehr et al.[31] Would we then be entitled to claim that the subjects all have a strong liking for reciprocity built into their preferences? I think the answer is no, because it is enough that some small percentage of subjects have such preferences for Nash equilibria to appear in which the employers make *high* offers and the workers put in *high* effort until the last few periods of the game.

We assume that each player is independently strategic with probability $1 - \pi$, or a reciprocating robot with probability π. A reciprocating robot makes a *high* offer as an employer and puts in *high* effort when receiving a *high* wage as a worker – until he observes that anyone at all has deviated from this behavior, after which he always plays *low*.

The following values of the parameters of the model admit a Nash equilibrium in which everybody plays *high* until the final period: $s = 1/10$, $b = 9/10$, $m = 8$, $n = 9$, and $\pi = 1/8$. The strategic workers mimic the programming of the robot workers until the last period, when they all play *low*. The strategic employers always play *high* until the penultimate period unless another employer publishes a *low* offer.[32] In the last two periods, the strategic employers mimic the programming of the robot employers. The Nash equilibrium described here is only the simplest of many with similar properties. I have not explored the full set of equilibria.

Strong reciprocity?

Herb Gintis is seconded by Paul Seabright in challenging the relevance of the folk theorem of repeated games to human cooperation. As well as Hobbists such as myself, they therefore take on evolutionary biologists such as Robert Trivers,[33] who invented the term 'reciprocal altruism' to capture the idea that the Nash equilibria which sustain cooperation in indefinitely repeated games operate on the basis of an 'I'll-scratch-your-back-if-you'll-scratch-mine' understanding.

Appealing to experimental results such as those of Fehr and his collaborators mentioned above, Gintis and Seabright argue against the orthodox notion of 'weak reciprocity' built into the folk theorem. They favor a notion of 'strong reciprocity', according to which evolution has wired a bias in favor of reciprocating into our preferences. That is to say, we reciprocate favors simply because we are programmed to like reciprocating favors.

Gintis gives two arguments for why we should reject the relevance of the folk theorem to human sociality, but before I address these, let me observe that he would seem thereby to undercut his own position. In criticizing Gintis on the same point, Ross argues that Gintis's story demands *discontinuous* changes in the evolutionary record. That is to say, the story would seem to require the appearance of what evolutionary biologists call 'hopeful monsters'. These are mutations responsible for *large* favorable changes in a mutant animal.

In my own putative evolutionary history for the device of the Rawlsian original position as the deep structure of fairness, I take great pains to argue that I am not relying on such hopeful monsters, because I share the orthodox view in evolutionary biology that their likelihood is so small that any story that depends on their emergence can be discarded as too fanciful for serious consideration. However, if Gintis were willing to allow weak reciprocity as a stepping stone toward strong reciprocity, it seems to me that he would no longer need to postulate the kind of hopeful monster that Ross criticizes. I do not think the experimental evidence supports the claim that human beings are as inflexible in their behavior as strong reciprocity would predict, but strong reciprocity would at least become more plausible than the usual kind of just-so evolutionary story of which social scientists are so fond.

Dynamic equilibrium?

Of the two objections that Gintis raises to the folk theorem, the first is that the idea of a Nash equilibrium should be replaced by his notion of a dynamic equilibrium. This would certainly sink the whole idea, since no pure strategy in an indefinitely repeated game can be a dynamic equilibrium for the same reason that no single pure strategy can be an evolutionarily, stable strategy according to the formal definition of Maynard Smith and Price.[34] But what is evolution then supposed to do in a repeated game – pack up and go home?

Larry Samuelson and I do not think so.[35] We argue that evolution will lead to a population *drifting* for long periods of time through whole sets of payoff-equivalent Nash equilibria of a repeated game. Even if we are only halfway right, accepting Gintis's proposal would then be a large step in entirely the wrong direction.

Expensive policing?

The second objection Gintis raises to the folk theorem relates to the cost of monitoring and punishing deviations from cooperative behavior. I will not treat this point at length, because Ross has done so already.

We think that monitoring your neighbors' observance of a social norm is largely costless in a small, close-knit society, because you have to monitor the behavior of those with whom you interact on a daily basis anyway. My own experience of living in a small village is that, far from finding it costly, my neighbors actively enjoyed keeping tabs on my eccentric comings and goings and those of my weirdo friends and relations. Nor need punishment often be costly if potential deviants can be kept under control most of the time by the kind of slowly increasing sanctions that hunter-gatherer groups are reported to employ.

Even when monitoring and punishment are costly, there is the consideration that if you do not carry out this part of your duties according to the reigning social contract, then you risk becoming the target of social disapproval yourself. When Gintis denies that such 'second-order punishment' is part of the human experience, one has to wonder on what planet he has been living all these years. If there is one safe thing to say in social science, it is surely that an insider who treats an outsider as an insider risks losing his own insider status.

Horns of a dilemma?

Paul Seabright argues that I am caught on the horns of a dilemma that derives from the fact that self-interested bargaining will serve as an equally good substitute for a fairness norm if we are looking for a device to solve equilibrium selection problems. It would indeed – moreover, it frequently does replace appeals to fairness in modern commercial life, which is hard to explain if you are committed to the view that we have other-regarding preferences wired into our genes.

However, I do not argue that fairness as an equilibrium selection device evolved under the conditions of modern commercial life, nor yet under the conditions of modern hunter-gatherer societies. I argue that our capacity to solve coordination problems by appealing to fairness criteria is part of what separates us from other animals. It therefore evolved before we were human, and certainly before language had developed to a stage at which it would have been possible to conduct a 'negotiation', as we understand the term in modern times. As I say in Chapter 2 of *Natural Justice*, we had to have some way of coordinating with each other to create the conditions under which true language could evolve.[36]

I go on to say that we mostly use fairness norms nowadays in situations in which it is either impossible to negotiate directly or in which the benefits of negotiating are outweighed by the costs in time or money. But I am not one of those right-wing economists who argue that fairness is a social fossil that should be discarded insofar as possible at the earliest opportunity. I think we should seek to understand how it works now in small-scale coordination problems, with a view to making use of this convenient piece of evolutionary flotsam to solve the large-scale coordination problems in the face of which direct negotiation seems helpless.

Conclusion

I know that what the Butlerians say is attractive – we all want to believe that people are fundamentally nice. If we had the choice, nobody would choose to live in a Hobbist world in which we need always to be on the alert against those who lie, cheat, and steal.

This is why Kant invented his categorical imperative. This is why so much energy was spent in the past inventing fallacious proofs that it is rational to cooperate in the prisoner's dilemma.[37] This is why Axelrod[38] stopped running his computer simulations before the nasty machines that evolution generates in the long run appeared.[39] This is why Edward Wilson was assaulted and denied a platform when seeking to communicate his sociobiological ideas.[40] This is why Butlerians do not look beyond the naive interpretation they offer of the results from their restricted menu of games; why they gloss over the results from mainstream experimental economics; and why they show no interest in the psychological literature on fairness and equity.[41]

The essence of a scientific approach is that we be willing to face up to unpleasant facts – that we recognize that all that glitters is not gold. Those of us who are realistic about what evolution has made of human nature have some chance of initiating reforms that might actually work, but all we can expect from the failure of the utopias invented by those who prefer to live in cloud cuckoo land is that we end up worse off than before. But nobody contributing to this issue merits the scorn that Aristophanes directed at Socrates and his impractical followers. In spite of our family differences, all four of us are brothers under our very thick skins. We do not agree on what evolution has made of human nature, but we are all committed to the unpopular task of trying to treat the question scientifically.

notes

[1] Ken Binmore, *Natural Justice* (Oxford: Oxford University Press, 2005).

[2] D. Munro, *A Guide to the British Moralists* (London: Collins, 1972).

[3] R. Frank, *Passions within Reason* (New York: Norton, 1968).

[4] D. Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986).

[5] C. Darwin, *The Expression of the Emotions in Man and Animals* (Chicago, IL: University of Chicago Press, 1965).

[6] K. Binmore, *Game Theory and the Social Contract, Volume 1: Playing Fair* (Cambridge, MA: MIT Press, 1994).

[7] *USA Today*, 8 April 2004.

[8] Joseph Henrich, 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* (New York: Oxford University Press, 2004).

[9] Ibid., p. 376.

[10] J. Henrich et al., '"Economic Man" in Cross-Cultural Perspective', *Behavioral and Brain Sciences* (2005).

[11] There is a risk of confusion when the repeated play of a one-shot game is under discussion. The assumption is then that players never expect to interact with their current opponents again. Unlike the repeated games to which the folk theorem applies, selfish optimizers will then have no reason to worry about being punished tomorrow for failing to cooperate today.

[12] J. Ledyard, 'Public Goods: A Survey of Experimental Research', in *Handbook of Experimental Game Theory*, edited by J. Kagel and A. Roth (Princeton, NJ: Princeton University Press, 1995); D. Sally, 'Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992', *Rationality and Society* 7 (1995): 58–92.

[13] C. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton, NJ: Princeton University Press, 2003).

[14] Henrich et al., *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, p. 433.

[15] E. Fehr and K. Schmidt, 'A Theory of Fairness, Competition and Cooperation', *Quarterly Journal of Economics* 114 (1999): 817–68.

[16] A. Shaked, 'The Rhetoric of Inequity Aversion', URL (2005): http://www.wiwwi.uni-bonn.de/shaked/rhetoric.

[17] Fehr and Schmidt tell us in their reply to Shaked's critique that they 'picked the value $\beta_i = .6$' on the grounds that it 'seemed more plausible to us'. See E. Fehr and K. Schmidt, 'The Rhetoric of Inequity Aversion – A Reply', URL (2005): http://www.wiwwi.uni-bonn.de/shaked/rhetoric. It certainly fits the data they claim to predict better than other values in the range between 0.5 and 1, which is all the accuracy their 'calibration' allows.

[18] K. Binmore, J. McCarthy, G. Ponti, L. Samuelson and A. Shaked, 'A Backward Induction Experiment', *Journal of Economic Theory* 104 (2002): 48–88.

[19] I say 'retro-classical' because Butlerians have abandoned the theory of revealed preference in favor of the Victorian view that people really do have utility functions in their heads.

20. K. Binmore, J. Gale and L. Samuelson, ‘Learning to be Imperfect: The Ultimatum Game’, *Games and Economic Behavior* 8 (1995): 56–90.
21. Why does evolution not eliminate responders who punish deviant proposers by saying ‘no’ when they are offered something positive, but ‘unfair’? Because, near the equilibrium, proposers deviate so little that the evolutionary pressures against such responders are negligible. See ibid.
22. E. Fehr and S. Gächter, ‘Cooperation and Punishment in Public Goods Experiments’, *American Economic Review* 90 (2000): 980–94.
23. Learning is no problem for Butlerians when people learn to play as Butlerians would wish.
24. J. Steiner, ‘A Trace of Anger is Enough: On the Enforcement of Social Norms’, CERGE-EI working paper (Prague: CERGE-EI, 2004).
25. E. Fehr, S. Gächter and G. Kirchsteiger, ‘Reciprocity as a Contract Enforcement Device: Experimental Evidence’, *Econometrica* 65 (1997): 833–60.
26. Ibid.
27. P.J. Healy, ‘Group Reputations and Stereotypes as Contract Enforcement Devices’, Caltech working paper (Pasadena, CA: Caltech, 2004) reports similar final-round effects, not only in his own instructive gift-exchange experiment, but also in those of M. Rigdon, ‘Efficiency Wages in an Experimental Labor Market’, *Proceedings of the National Academy of Sciences* 99 (2002): 13,348–51 and A. Riedl and J-R. Tyran, ‘Tax Liability Side Equivalence in Gift-Exchange Labor Markets’, *Journal of Public Economics* (2005 forthcoming).
28. M. Kandori, ‘Social Norms and Community Enforcement’, *Review of Economic Studies* 59 (1992): 63–80.
29. R. Selten and R. Stocker, ‘End Behavior in Finite Sequences of Prisoner’s Dilemma Supergames: A Learning Theory Approach’, *Journal of Economic Behavior and Organization* 7 (1986): 47–70.
30. This need not be attributed to a genetic disposition to reciprocate. I think such folk are simply more heavily conditioned to observe the social norms of everyday life than the rest of us.
31. E. Fehr et al., ‘Reciprocity as a Contract Enforcement Device: Experimental Evidence’.
32. Such an employer will be a reciprocating robot whose *high* wage was not reciprocated with *high* effort.
33. R. Trivers, ‘The Evolution of Reciprocal Altruism’, *Quarterly Review of Biology* 46 (1971): 35–56; R. Trivers, *Social Evolution* (Menlo Park, CA: Benjamin Cummings, 1985).
34. J. Maynard Smith and G. Price, ‘The Logic of Animal Conflict’, *Nature* 246 (1972): 15–18.
35. K. Binmore and L. Samuelson, ‘Evolutionary Stability in Repeated Games Played by Finite Automata’, *Journal of Economic Theory* 57 (1992): 278–305.
36. I do not argue that the use of the original position requires Adam and Eve to imagine bargaining directly behind the veil of ignorance, but that evolution will generate an outcome *as if* they had bargained directly.
37. Binmore, *Game Theory and the Social Contract, Volume 1: Playing Fair.*
38. R. Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).

39. K. Binmore, *Game Theory and the Social Contract, Volume 2: Just Playing* (Cambridge, MA: MIT Press, 1998).
40. J. Alcock, *The Triumph of Sociobiology* (New York: Oxford University Press, 2001).
41. G. Wagstaff, *An Integrated Psychological and Philosophical Approach to Justice* (Lampeter: Edwin Mellen Press, 2001).