KEYNES LECTURE

The Origins of Fair Play

KEN BINMORE
Fellow of the Academy

> All animals are equal but some are more equal than others.
>
> Orwell's *Animal Farm*

1. Introduction

THIS LECTURE IS A BRIEF OVERVIEW of an evolutionary theory of fairness. The ideas are fleshed out in a book *Natural Justice,* which is itself a condensed version of an earlier two-volume book *Game Theory and the Social Contract* (Binmore 2005, 1994, 1998).

2. How and why did fairness norms evolve?

My answer to the question *why?* is relatively uncontroversial among anthropologists. Sharing food makes good evolutionary sense, because animals who share food thereby insure themselves against hunger. It is for this reason that sharing food is thought to be so common in the natural world.

The vampire bat is a particularly exotic example of a food-sharing species. The bats roost in caves in large numbers during the day. At night, they forage for prey, from whom they suck blood if they can, but they are

Read at the Academy 28 February 2006.

Proceedings of the British Academy **151**, 159–193. © The British Academy 2007.

not always successful. If they fail to obtain blood for several successive nights, they die. The evolutionary pressure to share blood is therefore strong.

The biologist Wilkinson (1984) reports that a hungry bat begs for blood from a roostmate, who will sometimes respond by regurgitating some of the blood it is carrying in its own stomach. This is not too surprising when the roostmates are related, but the bats also share blood with roostmates who are not relatives. The behaviour is nevertheless evolutionarily stable, because the sharing is done on a *reciprocal* basis, which means that a bat is much more likely to help out a roostmate that has helped it out in the past. Bats that refuse to help out their fellows therefore risk not being helped out themselves in the future.

Vampire bats have their own way of sharing, and we have ours. We call our way of sharing 'fairness'. If the accidents of our evolutionary history had led to our sharing in some other way, it would not occur to us to attribute some special role to our current fairness norms. Whatever alternative norms we then found ourselves using would seem as solidly founded as those we find ourselves using today.

The *how?* questions are more troublesome. How do our current fairness norms work? How did they evolve? Both questions need to be addressed together, because each throws light on the other. In particular, I think that we need to be sceptical about answers to the first *how?* question that require our postulating 'hopeful monsters' when we seek to answer the second *how?* question. Richard Dawkins (1976) tirelessly explains how the eye might have evolved as the end-product of a process involving many small steps. We need to be able to do the same for the evolutionary processes that created our sense of fairness.

3. The Original Position

How do our fairness norms work? My thesis is that all the fairness norms that we actually use in daily life have a common deep structure that is captured in a stylised form by an idea that John Rawls (1972) called the device of the *original position* in his celebrated *Theory of Justice.*

Rawls—who is commonly said to be the leading moral philosopher of the last century—uses the original position as a hypothetical standpoint from which to make judgements about how a just society would be organised. Members of a society are asked to envisage the social contract to which they would agree *if* their current roles were concealed from them behind a ‘veil of ignorance’. Behind this veil of ignorance, the distribution of advantage in the planned society would seem determined as though by a lottery. Devil take the hindmost then becomes an unattractive principle for those bargaining in the original position, since you yourself might end up with the lottery ticket that assigns you to the rear.

Rawls defends the device of the original position as an operationalisation of Immanuel Kant’s categorical imperative, but I think this is just window-dressing. The idea certainly hits the spot with most people when they hear it for the first time, but I do not believe this is because they have a natural bent for metaphysics. I think it is because they recognise a principle that matches up with the fairness norms that they actually use every day in solving the equilibrium selection problem in the myriads of small coordination games of which daily life largely consists.

It is important to emphasise that I am not following Rawls here in talking about the major coordination problems faced by a nation state. Our sense of fairness did not evolve for use on such a grand scale. Nor am I talking about the artificial and unrealistic principles of justice promoted by self-appointed moral pundits, to which people commonly offer only lip service. Nor am I talking about the kind of moral pathology that led Osama bin Laden to believe that thousands of innocent New Yorkers should die to compensate for the humiliations that he thought Islam had received at the hands of the West. I am talking about the real principles that we actually use in solving everyday coordination problems.

The sort of coordination problems I have in mind are those that we commonly solve without thought or discussion, usually so smoothly and effortlessly that we do not even notice that there is a coordination problem to be solved. Who goes through that door first? How long does Adam get to speak before it is Eve’s turn? Who moves how much in a narrow corridor when a fat lady burdened with shopping passes a teenage boy with a ring through his nose? Who should take how much of a popular dish of which there is not enough to go around? Who gives way to whom when cars are manoeuvring in heavy traffic? Who gets that parking space? Whose turn is it to wash the dishes tonight? These are picayune problems, but if conflict arose every time they needed to be solved, our societies would fall apart.

Most people are surprised at the suggestion that there might be something problematic about how two people pass each other in the corridor. When interacting with people from our own culture, we commonly solve such coordination problems so effortlessly that we do not even think of them as problems. Our fairness programme then runs well below the level of consciousness, like our internal routines for driving cars or tying shoelaces. As with Molière's Monsieur Jourdain, who was delighted to discover that he had been speaking prose all his life, we are moral in small-scale situations without knowing that we are moral.

Just as we only take note of a thumb when it is sore, we tend to notice moral rules only when attempts are made to apply them in situations for which they are ill-adapted. We are then in the same position as Konrad Lorenz (1997) when he observed a totally inexperienced baby jackdaw go through all the motions of taking a bath when placed on a marble-topped table. By triggering such instinctive behaviour under pathological circumstances, Lorenz learned a great deal about what is instinctive and what is not when a bird takes a bath. But this vital information is gained only by avoiding the mistake of supposing that bath-taking behaviour confers some evolutionary advantage on birds placed on marble-topped tables.

Similarly, one can learn a lot about the mechanics of moral algorithms by triggering them under pathological circumstances—but only if one does not make the mistake of supposing that the moral rules are adapted to the coordination problems they fail to solve. However, it is precisely from such sore-thumb situations that I think traditional moralists unconsciously distil their ethical principles. We discuss these and only these situations endlessly, because our failure to coordinate successfully brings them forcefully to our attention.

This is not to say that we should not talk about such games. On the contrary, it is partly because we need to extend the class of games that our social contract handles adequately that it is worth studying the problem at all. But we will not learn how natural morality works by confining our attention to situations where it does not.

4. Justice as fairness

John Rawls (1972) offers a theory that reduces our notions of justice to those of fairness. I think our traditonal personification of justice as a blindfolded maiden bearing a pair of scales in one hand and a sword in the other provides some support for this reduction. Her blindfold can be identified with Rawls's veil of ignorance. She needs her scales behind the veil of ignorance to weigh up the relative well-being of different people in different situations.

The issue of how interpersonal comparisons are to be made is often treated as a side issue of no great importance by tradititional moral philosophers, but it is clearly necessary for people to be able to make such comparisons in order for it to be possible for them to use the original position to make fairness judgements. If we were not able to say whether we thought it preferable to be Adam in one situtation as opposed to being Eve in another situation, we would be helpless to say anything at all behind the veil of ignorance. Under mild conditions, John Harsanyi (1977) showed that such empathetic preferences—preferences requiring us to put ourselves in the position of another to see things from their point of view—can be summarised by naming a rate at which Adam's units of utility are to be traded off against Eve's units. But how do we acquire such standards of interpersonal comparison to which we implicitly appeal every time we make a fairness judgement?

Finally, attention needs to be drawn to the sword carried by our blindfolded maiden. The enforcement question is often neglected altogether by traditional moral philsophers, who commonly take for granted that fairness exists to trump the unbridled use of power that they think would otherwise reign supreme. However, I shall be arguing that fairness evolved as a means of *balancing* power rather than as a *substitute* for power. Without power being somehow exercised in her support, our blindfolded maiden would be no more than a utopian fancy. As Thomas Hobbes put it: 'Covenants without the sword are but words.'

5. Choosing between traditions

When I argue against traditional moral philosophers, I have in mind the metaphysical tradition that begins with Plato, and continues through Descartes and Kant to modern times, where it is firmly established as the reigning orthodoxy. Even John Harsanyi and John Rawls, from whom I draw much of my inspiration, regarded themselves as Kantians.

However, the naturalistic tradition is just as venerable. It begins with Aristotle, and continues through Epicurus, Hobbes and Hume, to the present day. Its leading modern exponent was John Mackie (1977), whose *Ethics: Inventing Right and Wrong* seems to me to offer a devastating critique of the orthodox view that morality somehow has an absolute status unconnected with the biological and social history of the human species. Instead of imagining that it is adequate in studying human morality to adopt the pose of Rodin's *Thinker* and await inspiration, he tells us to read the works of anthropologists and game theorists.

It is to this project that this paper and the books from which it is derived are devoted. It is particularly important to understand that the project requires disavowing Immanuel Kant on moral questions. If his categorical imperative implies anything specific, it surely calls for cooperation in the one-shot Prisoners' Dilemma, but his claim that such behaviour is *rational* seems absurd to game theorists. Our philosophical hero is David Hume, who was preaching our creed to an uncomprehending audience two hundred years before the first game theorist was born.

6. Pure foraging societies

There is no shortage of cultural differences between Kalahari bushmen, African pygmies, Andaman islanders, Greenland eskimos, Australian aborigines, Paraguayan Indians, and Siberian nomads, but the consensus is strong among modern anthropologists that these and other pure hunter-gatherer societies that survived into the twentieth century all operated social contracts without bosses or social distinctions in which food, especially meat, was shared on a markedly egalitarian basis.[1] Even Westermarck, a leading anthroplogist who was famous for his moral relativism, agreed that the Golden Rule—that we should do as we would be done by—was universally endorsed in such societies.

Two caveats are important here. The first is that it really matters that we are talking about *pure* foraging societies, in which the economic means of production remained the same as among our ancestors before the agricultural revolution of ten thousand or more years ago. The evidence is strong that a society's social contract evolves in tandem with its economy. I suspect that one would look in vain for universal principles underlying the social contracts that cultural history generated in different times and places after the agricultural revolution.

The second caveat is that one needs to put aside the idea that the egalitarianism of pure foraging societies makes them pastoral idylls, inhabited by noble savages filled with sweetness and light. Infanticide and murder are common. So is selfishness. Citizens of foraging societies do not honour their social contract because they like giving up food when they are

[1] See, for example, Bailey (1991), Damas (1972), Erdal and Whiten (1996), Evans-Pritchard (1940), Fürer-Haimendorf (1967), Gardner (1972), Hawkes *et al.* (1993), Helm (1972), Isaac (1978), Kaplan and Hill (1985), Knauft (1991), Lee (1979), Riches (1982), Tanaka (1980), Megarry (1995), Meggitt (1962), Rogers (1972), Sahlins (1974), Turnbull (1965).

hungry. They will therefore cheat on the social contract by secretly hoarding food if they think they can get away with it. The reason they comply with the norm most of the time is because their fellows will punish them if they do not.

Nor is there necessarily anything very nice about the way that food and other possessions are shared. In some societies, a fair allocation is achieved through 'tolerated stealing'. Eve may grab some of Adam's food because she thinks he has more than his fair share. If the rest of the group agree, Adam is helpless to resist. Even when possessions are voluntarily surrendered to others, the giver will sometimes explain that he or she is only complying with the norm to avoid being the object of the envy that precedes more serious sanctions. Indeed, we would find it unbearably stifling to live in some foraging societies because of the continual envious monitoring of who has what.

There is therefore squabbling and pettiness aplenty in pure foraging communities, but there is also laughter and good fellowship. In brief, human nature seems much the same in foraging societies as in our own. I therefore think the strong parallels that anthropologists have uncovered between the social contracts of geographically distant groups living in starkly different environments have important implications for us. If their nature includes an instinctive disposition to use fairness norms that all share the deep structure of the Rawlsian original position, is it not likely that the same disposition is built into our nature too?

7. Game theory

John Mackie invited us to look at both anthropology and game theory. The basic idea in game theory is that of a Nash equilibrium. John Nash was the subject of the movie *A Beautiful Mind,* but the writers of the movie got the idea hopelessly wrong in the scene where they tried to explain how Nash equilibria work. However, the idea is actually very simple.

A game is any situation in which people or animals interact. The plans of action of the players are called strategies. A Nash equilibrium is any profile of strategies—one for each player—in which each player's strategy is a best reply to the strategies of the other players.

Some simple examples appear in Figure 1. The game on the left is the famous Prisoners' Dilemma. The game on the right is the Stag Hunt, which game theorists use to illustrate a story of Jean-Jacques Rousseau.

	dove	*hawk*
dove	2 2	3* 0
hawk	0 3*	1* 1*

Prisoners' Dilemma

	dove	*hawk*
dove	4* 4*	3 0
hawk	0 3	2* 2*

Stag Hunt Game

Figure 1. Two toy games.

Each of these toy games has two players, whom I call Adam and Eve. In both the Prisoners' Dilemma and the Stag Hunt, Adam has two strategies, *dove* and *hawk*, that are represented by the rows of the payoff table. Eve also has two strategies, *dove* and *hawk*, represented by the columns of the payoff table. The four cells of the payoff table correspond to the possible outcomes of the game. Each cell contains two numbers, one for Adam and one for Eve. The number in the south-west corner is Adam's payoff for the corresponding outcome of the game. The number in the north-east corner is Eve's payoff.

The payoffs will not usually correspond to money in the applications relevant in this lecture. Using the theory of revealed preference, economists have shown that *any* consistent behaviour whatever can be modelled by assuming that the players are behaving as though seeking to maximise the average value of *something*. This abstract something—which obviously varies with the context—is called utility. When assuming that a player is maximising his or her expected payoff in a game, we are therefore not taking for granted that people are selfish. In fact, we make no assumptions about their motivation except that the players pursue their goals—whatever they may be—in a consistent manner.

It would be easy for the players to maximise their expected payoffs if they knew what strategy their opponent was going to choose. For example, if Adam knew that Eve was going to choose *dove* in the Prisoners' Dilemma, he would maximise his payoff by choosing *hawk*. That is to say, *hawk* is Adam's best reply to Eve's choice of *dove*, a fact indicated in Figure 1 by starring Adam's payoff in the cell that results if the players choose the strategy profile (*hawk*, *dove*). However, the problem in game theory is that a player does not normally know in advance what strategy the other player will choose.

A Nash equilibrium is a strategy profile in which each player's strategy is a best reply to the strategies chosen by the other players. In the examples of Figure 1, a cell in which both payoffs are starred therefore corresponds to a Nash equilibrium.

Nash equilibria are of interest for two reasons. If it is possible to single out the rational solution of a game, it must be a Nash equilibrium. For example, if Adam knows that Eve is rational, he would be stupid not to make the best reply to what he knows is her rational choice. The second reason is even more important. An evolutionary process that adjusts the players' strategy choices in the direction of increasing payoffs can only stop when it reaches a Nash equilibrium.

Because evolution stops working at an equilibrium, biologists say that Nash equilibria are evolutionarily stable.[2] Each relevant locus on a chromosome is then occupied by the gene with maximal fitness. Since a gene is just a molecule, it cannot *choose* to maximise its fitness, but evolution makes it seem as though it had. This is a valuable insight, because it allows biologists to use the rational interpretation of an equilibrium to predict the outcome of an evolutionary process, without following each complicated twist and turn that the process might take.

The title of Richard Dawkins' (1976) *Selfish Gene* expresses the idea in a nutshell, but it also provokes a lot of criticism. It is easy to be tolerant of critics like the old lady I heard rebuking Dawkins for failing to see that a molecule can't possibly have free will, but tolerance is less easy in the case of critics like Lewontin or Gould, who chose to whip up public hostility against Edward Wilson and his followers on similar grounds. As Alcock's (2001) *Triumph of Sociobiology* documents, they wilfully pretended not to understand that sociobiologists seek explanations of biological phenomena in terms of *ultimate* causes rather than *proximate* causes.

Why, for example, do songbirds sing in the early spring? The proximate cause is long and difficult. This molecule knocked against that molecule. This chemical reaction is catalysed by that enzyme. But the ultimate cause is that the birds are signalling territorial claims to each other in order to avoid unnecessary conflict. They neither know nor care that this behaviour is rational. They just do what they do. But the net effect of an immensely complicated evolutionary process is that songbirds behave *as*

[2] John Maynard Smith (1982) defines an evolutionarily stable strategy (ESS) to be a best reply to itself that is a better reply to any alternative best reply than the alternative best reply is to itself, but biologists seldom worry about the small print involving alternative best replies.

though they had rationally chosen to maximise their fitness by operating a Nash equilibrium of their game of life.

The Prisoners' Dilemma is the most famous of all toy games. A whole generation of scholars swallowed the line that this trivial game embodies the essence of the problem of human cooperation. The reason is that its only Nash equilibrium calls for both Adam and Eve to play *hawk*, but they would both get more if they cooperated by both playing *dove* instead. The hopeless task that scholars set themselves was therefore to give reasons why game theory's resolution of this supposed 'paradox of rationality' is mistaken.

Game theorists think it just plain wrong to claim that the Prisoners' Dilemma embodies the essence of the problem of human cooperation. On the contrary, it represents a situation in which the dice are as loaded against the emergence of cooperation as they could possibly be. If the great game of life played by the human species were the Prisoners' Dilemma, we would not have evolved as social animals! We therefore see no more need to solve some invented paradox of rationality than to explain why strong swimmers drown when thrown in a lake with their feet encased in concrete. No paradox of rationality exists. Rational players do not cooperate in the Prisoners' Dilemma, because the conditions necessary for rational cooperation are absent in this game.

Fortunately the paradox-of-rationality phase in the history of game theory is just about over. Insofar as they are remembered, the many fallacies that were invented in hopeless attempts to show that it is rational to cooperate in the Prisoners' Dilemma are now mostly quoted as entertaining examples of what psychologists call magical reasoning, in which logic is twisted to secure some desired outcome. The leading example remains Kant's claim that rationality demands obeying his categorical imperative. In the Prisoners' Dilemma, rational players would then all choose *dove*, because this is the strategy that would be best if everybody chose it.

The following argument is a knock-down refutation of this nonsense. So as not to beg any questions, we begin by asking where the payoff table that represents the players' preferences in the Prisoners' Dilemma comes from. The economists' answer is that we discover the players' preferences by observing the choices they make (or would make) when solving one-person decision problems.

Writing a larger payoff for Adam in the bottom-left cell of the payoff table of the Prisoners' Dilemma than in the top-left cell therefore means that Adam would choose *hawk* in the one-person decision problem that he would face if he knew in advance that Eve had chosen *dove*. Similarly, writing a larger payoff in the bottom-right cell means that Adam would choose *hawk* when faced with the one-person decision problem in which he knew in advance that Eve had chosen *hawk*.

The very definition of the game therefore says that *hawk* is Adam's best reply when he knows that Eve's choice is *dove*, and also when he knows her choice is *hawk*. So he does not need to know anything about Eve's actual choice to know his best reply to it. It is rational for him to play *hawk* whatever strategy she is planning to choose. Nobody ever denies this utterly trivial argument. Instead, one is told that it cannot be relevant to anything real, because it reduces the analysis of the Prisoners' Dilemma to a tautology. But who would say the same of 2+2=4?

In Rousseau's Stag Hunt story, Adam and Eve agree to cooperate in hunting a stag, but when they separate to put their plan into action, each may be tempted to abandon the joint enterprise by the prospect of bagging a hare for themselves. The starred payoffs in the payoff table show that there are two Nash equilibria in pure strategies, one in which the players cooperate by both playing *dove*, and one in which they defect by both playing *hawk*.We therefore have our first example of the equilibrium selection problem that will be our major preoccupation in the rest of this lecture.

If a society found itself at a social contract corresponding to the inefficient equilibrium in which everybody plays *hawk*, why would not they just agree to move to the efficient social contract in which everybody plays *dove*?

As the biologist Sewell-Wright explained, this may not be so easy if the task of moving from one equilibrium to another is left to evolution. But we are not animals who have to wait for the slow forces of evolution to take them to a new social contract. We can talk to each other and agree to alter the way we do things. But can we trust each other to keep any agreement we might make? The Stag Hunt is used by experts in international relations under the name of the Security Dilemma to draw attention to the problems that can arise even when the players are rational.

Suppose that Adam and Eve's current social contract in the Stag Hunt is the Nash equilibrium in which they both play *hawk*. However hard Adam seeks to persuade Eve that he plans to play *dove* in the future and so she should follow suit, she will remain unconvinced. The reason is that whatever Adam is actually planning to play, it is in his interests to persuade Eve to play *dove*. If he succeeds, he will get 4 rather than 0 if he is planning to play *dove*, and 3 rather than 2 if he is planning to play *hawk*. Rationality alone therefore does not allow Eve to deduce anything about his plan of action from what he says, because he is going to say the same thing no matter what his real plan may be! Adam may actually think that Eve is unlikely to be persuaded to switch from *hawk* and hence be planning to play *hawk* himself, yet still try to persuade her to play *dove*.

This Machiavellian story shows that attributing rationality to the players is not enough to resolve the equilibrium selection problem—even in a seemingly transparent case like the Stag Hunt. If Adam and Eve continue to play *hawk* in the Stag Hunt, they will regret their failure to coordinate on playing *dove*, but neither can be accused of being irrational, because both are doing as well as they can given the behaviour of their opponent.

The standard response is to ask why game theorists insist that it is irrational for people to trust one another. Would not Adam and Eve both be better off if both had more faith in each other's honesty? But nobody denies that Adam and Eve would be better off if they trusted each other, any more than anybody denies that they would be better off in the Prisoners' Dilemma if they were assigned new payoffs that made them care more about the welfare of their opponent. Nor do game theorists say it is irrational for people to trust each other. They only say that it is not rational to trust people without a good reason: that trust cannot be taken on trust. Who trusts a used-car dealer or a dean? What wife doesn't keep an eye on her husband? Who does not count their change?

The underlying point here is that those of us who would like society to move to what we think will be a better social contract just waste our time if we simply bleat that people should be more trusting or honest. We need to try and understand how and why it makes sense to be trusting or honest in some situations, but not in others. We can then hope to improve our social contract by doing what we can to promote the former situations at the expense of the latter.

8. Coordination Games

I think that fairness evolved as Nature's answer to the equilibrium selection problem in the human game of life. However, before I can elaborate on this idea, it is necessary to give examples of some toy games in which the equilibrium selection problem is even more pressing than in the Stag Hunt.

The game on the left of Figure 2 is a simplified version of the Driving Game that we play each morning when we get into our cars and drive to work. It does not matter on which of its two Nash equilibria a society coordinates, but it is obviously very important that we all coordinate on the same equilibrium. In Britain and Japan, the accidents of our social history have resulted in our all driving on the left. In the USA and France, everybody drives on the right. Culture can therefore be a significant factor in the way we solve an equilibrium selection problem. Indeed, it can be argued that a society's culture is nothing more than the set of conventions that it uses to solve equilibrium selection problems. A politically incorrect story accompanies the Battle of the Sexes on the right of Figure 2. Adam and Eve are on their honeymoon in New York City. At breakfast, they discussed whether to attend a boxing match or the ballet in the evening, but without reaching an agreement. During the day they got separated in the crowds, and they must now choose where to go in the evening independently.

The Battle of the Sexes has two Nash equilibria,[3] in one of which Adam and Eve both go to the boxing match, and one in which they both go the ballet. But, unlike the case of the Driving Game, it now matters to the players which equilibrium is chosen, because Adam prefers boxing to ballet, and Eve prefers ballet to boxing.

9. Reciprocity

I have already signalled my intention of modelling a social contract as the set of common understandings in a society that allows its citizens to coordinate on one of the many equilibria of their game of life. Game theorists think that only equilibria are viable in this role, because, when each citizen

	left	*right*
left	1* 1*	0 0
right	0 0	1* 1*

Driving Game

	box	*ball*
box	1* 2*	0 0
ball	0 0	2* 1*

Battle of the Sexes

Figure 2. Coordination games.

[3] Only pure strategies are considered here.

has independent goals which sometimes conflict with each other, only equilibria can survive in the absence of an external enforcement agency. In brief, only equilibria are self-policing.

The suggestion that a social contract is no more than a set of common understandings among players acting in their own enlightened self interest commonly gets a sceptical reaction. How can anything very sturdy be erected on such a flimsy foundation? Surely a solidly built structure like the modern state must be firmly based on a rock of moral certitude, and only anarchy can result if everybody just does what takes his fancy? As Gauthier (1986: 1) expresses it in denying Hume (1975: 280): 'Were duty no more than interest, morals would be superfluous.'

I believe such objections to be misconceived. Firstly, there are no rock-like moral certitudes that exist prior to society. To adopt a metaphor that sees such moral certitudes as foundation stones is therefore to construct a castle in the air. Society is more usefully seen as a dynamic organism, and the moral codes that regulate its internal affairs are the conventional understandings which ensure that its constituent parts operate smoothly together when it is in good health. Moreover, the origin of these moral codes is to be looked for in historical theories of biological, social, and political evolution, and not in the works of abstract thinkers no matter how intoxicating the wisdom they distil. Nor is it correct to say that anarchy will necessarily result if everybody 'just' does what he wants. A person would be stupid in seeking to achieve a certain end if he ignored the fact that what other people are doing is relevant to the means for achieving that end. Intelligent people will *coordinate* their efforts to achieve their individual goals without necessarily being compelled or coerced by real or imaginary bogeymen.

The extent to which simple implicit agreements to coordinate on an equilibrium can generate high levels of cooperation among populations of egoists is not something that is easy to appreciate in the abstract. That *reciprocity* is the secret has been repeatedly discovered, most recently by the political scientist Axelrod (1984) in the eighties and the biologist Trivers (1971) in the seventies. However, David Hume (1978: 521) had already put his finger on the relevant mechanism some 230 years before:

> I learn to do service to another, without bearing him any real kindness: because I foresee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or others. And accordingly, after I have serv'd him and he is in possession of the advantage arising from my action, he is induc'd to perform his part, as foreseeing the consequences of his refusal.

In spite of all the eighteenth-century sweetness and light, one should take special note of what Hume says about foreseeing the consequences of refusal. The point is that a failure to carry out your side of the arrangement will result in your being *punished.* The punishment may consist of no more than a refusal by the other party to deal with you in future. Or it may be that the punishment consists of having to endure the disapproval of those whose respect is necessary if you are to maintain your current status level in the community. However, nothing excludes more active forms of punishment. In particular, the punishment might be administered by the judiciary, if the services in question are the subject of a legal contract.

At first sight, this last observation seems to contradict the requirement that the conventional arrangements under study be *self-policing*. The appearance of a contradiction arises because one tends to think of the apparatus of the state as somehow existing independently of the game of life that people play. But the laws that societies make are not part of the rules of this game. One *cannot* break the rules of the game of life, but one certainly can break the laws that man invents.

Legal rules are no more than particularly well-codified conventions. And policemen, judges and public executioners do not exist outside society. Those charged with the duty of enforcing the laws that a society formally enacts are themselves only players in the game of life. However high-minded a society's officials may believe themselves to be, the fact is that society would cease to work in the long run if the duties assigned to them were incompatible with their own incentives. I am talking now about corruption. And here I do not have so much in mind the conscious form of corruption in which officials take straight bribes for services rendered. I have in mind the long-term and seemingly inevitable process by means of which bureaucracies gradually cease to operate in the interests of those they were designed to serve, and instead end up serving the interests of the bureaucrats themselves.

10. The folk theorem

Game theorists rediscovered Hume's insight that reciprocity is the mainspring of human sociality in the early fifties when characterising the outcomes that can be supported as equilibria in a repeated game. The result is known as the *folk theorem,* since it was formulated independently by several game theorists in the early fifties (Aumann and Maschler 1995).

The theorem tells us that external enforcement is unnecessary to make a collection of Mr Hydes cooperate like Dr Jekylls. It is only necessary that the players be sufficiently patient and that they know they are to interact together for the foreseeable future. The rest can be left to their enlightened self-interest, provided that they can all monitor each other's behaviour without too much effort—as, for example, must have been the case when we were all members of small hunter-gatherer communities.

What outcomes can be sustained as Nash equilibria when a one-shot game is repeated indefinitely often? The answer provided by the folk theorem is very reassuring. Any outcome whatever of the one-shot game—including all the outcomes that are not Nash equilibria of the one-shot game—can be sustained as Nash equilibria of the *repeated* game, provided that they award each player a payoff that is at least as large as the player's minimax payoff in the one-shot game.

The idea of the proof is absurdly simple. We first determine how the players have to cooperate to obtain a particular outcome. For example, in the repeated Prisoners' Dilemma, Adam and Eve need only play *dove* at each repetition to obtain an average payoff of 2. In the repeated Battle of the Sexes, Adam and Eve can each get an average payoff of 1½ if they attend the boxing match on odd days and the ballet on even days. To make such cooperative behaviour into a Nash equilibrium, it is necessary that a player who deviates be punished. This is where the players' minimax payoffs enter the picture. The worst punishment that Eve can inflict on Adam in the long run is to hold him to his minimax payoff, because when she acts to try and *minimise* his payoff, he will respond by playing whatever strategy *maximises* his payoff given her choice of punishment strategy. As long as the average payoff a player gets by cooperating exceeds his minimax payoff, the player can therefore be kept in line if he knows that his opponents will respond to any deviation on his part by holding him to his minimax strategy.

In the Prisoners' Dilemma, the minimax payoff for each player is 1, because the worst that one player can do to the other is play *hawk*, in which case the victim does best to respond by playing *hawk* as well. The folk theorem therefore tells us that we can sustain the outcome in which both players always play *dove* as a Nash equilibrium in the indefinitely repeated game. In the Battle of the Sexes, the minimax payoff for each player is also 1.[4] For example, the worst that Eve can do to Adam is play

[4] Provided that we neglect the possible use of mixed strategies.

ball, in which case he does best to respond by playing *ball* as well. The folk theorem therefore tells us that we can sustain the outcome in which both players alternate between attending the boxing match and the ballet as an equilibrium in the repeated game.

The kind of equilibrium strategy described above is often called the *grim* strategy, because the punishment that keeps potential deviants on the straight and narrow path is the worst possible punishment indefinitely prolonged. One sometimes sees this strategy in use in commercial contexts where maintaining trust is vital to the operation of a market. The Antwerp diamond market is a good example. Traders pass diamonds back and forward for examination without any writing of contracts or attempt to monitor those trusted with diamonds on approval. Why do not the traders cheat each other? Because any suspicion of misconduct will result in a trader being excluded from the market thereafter. To quote a trader in the similar New York antique market: 'Sure I trust him. You know the ones to trust in this business. The ones who betray you, bye-bye' (*New York Times,* 29 August 1991).

However, one seldom sees the grim strategy in use in games played among social insiders. The punishments are then typically minimal rather than maximal, because the deviant requiring punishment may then turn out to be yourself, or one of your friends or relations. Napoleon's exile in Elba is an extreme example. After all, any ruler may be overthrown. On the other hand, we bourgeois folk do not ever expect to steal a pizza, and hence the Californian doctrine of three strikes and you are out.

It is important to recognise that very few of the punishments that sustain a social contract are administered through the legal system. Indeed, nearly all punishments are adminstered without either the punishers or the victim being aware that a punishment has taken place. No stick is commonly flourished. What happens most of the time is that the carrot is withdrawn a tiny bit. Shoulders are turned slightly away. Greetings are imperceptibly gruffer. Eyes wander elsewhere. These are all warnings that your body ignores at its peril.

The accounts that anthropologists offer of the higher stages of punishment observed among pure hunter-gatherer societies are particularly telling, since they mirror so accurately similar phenomena that the academic world uses to keep rogue thinkers in line. First there is laughter. If this does not work—and who likes being laughed at—the next stage is boycotting. Nobody talks to the offending party, or refers to his research. Only the final stage is maximal: a persistent offender is expelled from the group, or is unable to get his work published.

Once the subtle nature of the web of reciprocal rewards and punishments that sustains a social contract has been appreciated, it becomes easier to understand why it is so hard to reform corrupt societies in which criminality has become socially acceptable. As the case of Prohibition shows, imposing the type of draconian penalty in which rednecks delight on the criminals unlucky enough to be caught is unlikely to be effective. The resulting disincentives will be almost certainly be inadequate, since the probability of any individual being unlucky is necessarily small when nearly everybody is guilty.

11. Selecting Equilibria

The study of Nash equilibria in repeated games offers us clues about how social contracts are sustained. I think that deontological philosophers derive their inspiration from focusing their attention largely on such stability questions. Conservative economists are led to similar positions by choosing to examine models that only have one equilibrium.

However, the folk theorem tells us that indefinitely repeated games—including those markets that are repeated on a daily basis—usually have very large numbers of equilibria amongst which a choice must somehow be made. The response that only efficient equilibria need be considered does not help with this equilbrium selection problem, because efficient equilibria are also usually present in large numbers.[5]

One way of selecting an equilibrium is to delegate the task to a leader or an elite, but our foraging ancestors had no leaders or elites. Some other equilibrium selection device was therefore necessary. Fairness is our name for the device that evolution came up with. Consequentialist philosophers commonly offer metaphysical explanations of why their own idiosyncratic theories of fairness should prevail, but the truth is simultaneously more complex and more prosaic. Our ancestors were fair for much the same reason that the French drive on the right and the Japanese on the left. Any solution to the equilibrium selection problem is better than none.

However, the consequentialists and the radical reformers of the left whom they inspire make a more serious mistake when they fail to appre-

[5] A (Pareto) efficient outcome is one on which no player can improve his payoff without making another player worse off. For example, the equilibria (*box*, *box*) and (*ball*, *ball*) are both efficient outcomes for the Battle of the Sexes.

ciate that fairness evolved to select among *equilibria*—that fairness norms that actually work are not substitutes for power, but merely help to determine how power is balanced. Deontologists and their conservative followers do not make this mistake. Instead, they close their eyes to the possibility of rational reform by failing to recognise that there may be alternative equilibria to those with which they familiar.

12. Deep structure of fairness

Recognition of the Golden Rule seems to be universal in human societies. Is there any reason why evolution should have written such a principle into our genes? Some equilibrium selection devices are obviously necessary for social life to be possible, but why should something like the Golden Rule have evolved?

If the Golden Rule is understood as a simplified version of the device of the original position, I think an answer to this question can be found by asking why social animals evolved in the first place. This is generally thought to have been because food-sharing has survival value.

The vampire bats of Section 2 provide an example. Unless a vampire bat can feed every sixty hours or so, it is likely to die. The advantages of sharing food among vampire bats are therefore strong—so strong that evolution has taught even unrelated bats to share blood on a reciprocal basis.

By sharing food, the bats are essentially *insuring* each other against hunger. Animals cannot write insurance contracts in the human manner, and even if they could, they would have no legal system to which to appeal if one animal were to hold up on his or her contractual obligation to the other. But the folk theorem tells us that evolution can get round the problem of external enforcement if the animals interact together on a *repeated* basis.

By coordinating on a suitable equilibrium in their repeated game of life, two animals who are able to monitor each other's behaviour sufficiently closely can achieve whatever could be achieved by negotiating a legally binding insurance contract. It will be easier for evolution to find its way to such an equilibrium if the animals are related, but the case of vampire bats shows that kinship is not necessary if the evolutionary pressures are sufficiently strong.

What considerations would Adam and Eve need to take into account when negotiating a similar mutual insurance pact?

Imagine a time before cooperative hunting had evolved, in which Adam and Eve foraged separately for food. Like vampire bats, they would sometimes come home lucky and sometimes unlucky. An insurance pact between them would specify how to share the available food on days when one was lucky and the other unlucky.

If Adam and Eve were rational players negotiating an insurance contract, they would not know in advance who was going to be lucky and who unlucky on any given day on which the contract would be invoked. To keep things simple, suppose that both possibilities are equally likely. Adam and Eve can then be seen as bargaining behind a *veil of uncertainty* that conceals who is going to turn out to be Ms Lucky or Mr Unlucky. Both players then bargain on the assumption that they are as likely to end up holding the share assigned to Mr Unlucky as they are to end up holding the share assigned to Ms Lucky.

I think the obvious parallel between bargaining over such mutual insurance pacts and bargaining in the original position is no accident. To nail the similarity down completely, we need only give Adam and Eve new names when they take their places behind Rawls's veil of ignorance. To honour the founders of game theory, Adam and Eve will be called John and Oskar.

Instead of Adam and Eve being uncertain about whether they will turn out to be Ms Lucky or Mr Unlucky, the new setup requires that John and Oskar pretend to be ignorant about whether they will turn out to be Adam or Eve. It then becomes clear that a move to the device of the original position requires only that the players imagine themselves in the shoes of somebody else—either Adam or Eve—rather than in the shoes of one of their own possible future selves.

If Nature wired us up to solve the simple insurance problems that arise in food-sharing, she therefore also simultaneously provided much of the wiring necessary to operate the original position.

Of course, in an insurance contract, the parties to the agreement do not have to *pretend* that they might end in somebody else's shoes. On the contrary, it is the reality of the prospect that they might turn out to be Ms Lucky or Mr Unlucky that motivates their writing a contract in the first place. But when the device of the original position is used to adjudicate fairness questions, then John knows perfectly well that he is actually Adam, and that it is physically impossible that he could become Eve. To use the device in the manner recommended by Rawls and Harsanyi, he therefore has to indulge in a counterfactual act of imagination. He cannot become Eve, but he must pretend that he could. How is this gap between reality and pretence to be bridged without violating the Linnaean dictum: *Natura non facit saltus*?

As argued earlier, I think that human ethics arose from Nature's attempt to solve certain equilibrium selection problems. But Nature does not jump from the simple to the complex in a single bound. She tinkers with existing structures rather than creating hopeful monsters. To make a naturalistic origin for the device of the original position plausible, it is therefore necessary to give some account of what tinkering she might have done.

In Peter Singer's *Expanding Circle* (1980), the circle that expands is the domain within which moral rules are understood to apply. For example, Jesus sought to expand the domain of the principle that you should love your neighbour by redefining a neighbour to be anyone at all. How might evolution expand the domain within which a moral rule operates?

My guess is that the domain of a moral rule sometimes expands when players misread signals from their environment, and so mistakenly apply a piece of behaviour or a way of thinking that has evolved for use within some inner circle to a larger set of people, or to a new game. When such a mistake is made, the players attempt to play their part in sustaining an equilibrium in the inner-circle game without fully appreciating that the outer-circle game has different rules. For example, Adam might treat Eve as a sibling even though they are unrelated. Or he might treat a one-shot game as though it were going to be repeated indefinitely often.

A strategy profile that is an equilibrium for an inner-circle game will not normally be an equilibrium for an outer-circle game. A rule that selects an equilibrium strategy in an inner-circle game will therefore normally be selected against if used in an outer-circle game. But there will be exceptions. When playing an outer-circle game as though it were an inner-circle game, the players will sometimes happen to coordinate on an equilibrium of the outer-circle game. The group will then have stumbled upon an equilibrium selection device for the outer-circle game. This device consists of the players behaving *as though* they were constrained by the rules of the inner-circle game, when the rules by which they are actually constrained are those of the outer-circle game.

I guess that nobody questions Aristotle's observation that the origins of moral behaviour are to be found in the family. A game theorist will offer the explanation that the equilibrium selection problem is easier for evolution to solve in such games. The reason why is to be found in Hamilton's (1963) rule, which explains that animals should be expected to care about a relative in proportion to their degree of relationship to the relative. For example, if Eve is Adam's full cousin, it makes evolutionary sense for him to count her fitness as worth one-eighth of his own fitness.[6] Family relationships therefore provide a natural basis for making the kind of interpersonal comparison of utility that is necessary to operate the device of the original position.

The circle was then ready to be expanded by including strangers in the game by treating them as honorary or fictive kinfolk, starting with outsiders adopted into the clan by marriage or cooption. Indeed, if you only interact on a regular basis with kinfolk, what other template for behaviour would be available?

The next step requires combining these two developments so that the original position gets to be used not just in situations in which Adam and Eve might turn out to be themselves in the role of Ms Lucky or Mr Unlucky, but in which they proceed as though it were possible for each of them to turn out to occupy the role of the other person. To accept that I may be unlucky may seem a long way from contemplating the possibility that I might become another person in another body, but is the difference really so great? After all, there is a sense in which none of us are the same person when comfortable and well fed as when tired and hungry. In different circumstances, we reveal different personalities and want different things.

To pursue this point, consider what is involved when rational players consider the various contingencies that may arise when planning ahead. To assess these, players compute their expected utility as a weighted average of the payoffs of all the future people—lucky or unlucky—that they might turn out to be after the dice has ceased to roll. When choosing a strategy in a family game, players similarly take their payoffs to be a weighted average of the fitnesses of everybody in their family.

In order to convert our ability to negotiate insurance contracts into a capacity for using fairness as a more general coordinating device in the game of life, all that is then needed is for us to hybridise these two processes by allowing players to replace one of the future persons that a roll of the dice might reveal them to be, by a person in another body. The empathetic preferences that are needed to assess this possibility require nothing more than that they treat this person in another body in much the same way that they treat their sisters, cousins or aunts.

[6] Because the probability that a newly mutated gene in his body responsible for modifying some relevant behaviour is also in her body is 1/8.

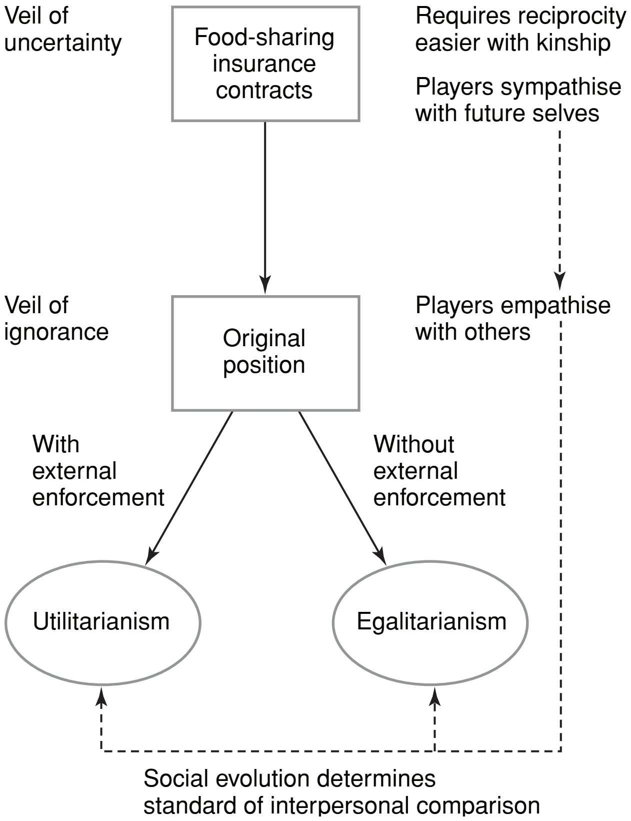
Veil of uncertainty
Food-sharing insurance contracts
Requires reciprocity easier with kinship
Players sympathise with future selves
Veil of ignorance
Original position
Players empathise with others
With external enforcement
Without external enforcement
Utilitarianism
Egalitarianism
Social evolution determines standard of interpersonal comparison

Figure 3. An evolutionary history of the original position?

Figure 3 illustrates the evolutionary history of the original position in the story told so far. It also draws attention to the need to consider the source of the standard of interpersonal comparison built into the empathetic preferences with which Adam and Eve enter the original position. I follow the psychology literature in specifying this standard by assigning positive numbers to Adam and Eve, but I refer to these positive numbers as social indices rather than worthiness coefficients.

The social indices we use when discounting the fitnesses of our partners in a family game are somehow obtained by estimating our degree of relationship to our kinfolk from the general dynamics of the family and our place in it. But where do we get the social indices with which to discount Adam and Eve's personal utils when constructing an empathetic utility function?

I do not think that we acquire the social indices we apply to different people at different times and in different contexts through any process of conscious reflection. Still less do we consult the works of moral philosophers. We pick up the appropriate social indices—in much the same way as we pick up most of our social behaviour—from unconsciously imitating the behaviour of those of our fellow citizens whom we admire or respect. That is to say, I attribute our standard of interpersonal comparison of utility in dealing with folk outside our intimate circle of family and friends to the workings of social or cultural evolution.

13. Enforcement

The previous section offers a putative evolutionary explanation for why we represent Justice as a blindfolded maiden bearing a pair of scales. But what of the sword that represents her powers of enforcement?

As Figure 3 indicates, this makes all the difference between whether the use of the original position leads to utilitarian or egalitarian conclusions. Space does not allow a review of the argument, but if we follow Harsanyi in assuming that the hypothetical deal reached in the original position is enforced by some outside agency, then the outcome will be utilitarian (Binmore 2005: chapter 10). On the other hand, if we admit no external enforcement at all, then the outcome is egalitarian (Binmore 2005: chapter 11).

Harsanyi (1977) invents an agency called 'moral commitment' that somehow enforces the hypothetical deal reached in the original position. Rawls (1972) similarly invents an agency called 'natural duty' for the same purpose. My own view is that we are not entitled to invent anything at all. If we treat the government of a modern state as an omnipotent but benign power whose function is to enforce the decisions made by the people under fair conditions, then Harsanyi's analysis provides a reason why the government should make decisions on a utilitarian basis. However, if there is no real (as opposed to invented) external enforcement agency, then Harsanyi's argument fails. In particular, it fails if the officers of a government are themselves treated as people with their own personal interests, just like any other citizen.

How come that Harsanyi is led to a utilitarian conclusion and Rawls to an egalitarian conclusion, given that they begin with the same assumptions? Game theorists trace the reason to Rawls's decision to deny orthodox decision theory. Without this iconoclastic expedient, he too would have been led to a utilitarian conclusion—although his *Theory of Justice* was explicitly written to provide a reasoned alternative to utilitarianism. My own view is that Rawls's purposes would have been better served if he had taken more seriously the concerns he refers to as the 'strains of commitment' in the third and longest part of his book. Taken to their logical extreme, these stability considerations require that everything involved in operating the original position must be self-policing. But then we are led to an egalitarian position not so very different from that he defends in his *Theory of Justice.*

There is, in fact, some empirical support for the kind of egalitarian sharing to which one is led by analysing the result of bargaining in the original position when all the arrangements must be self-policing.[7] As in Wilson (1983), the theory is usually called 'modern' equity theory, although it goes all the way back to Aristotle (1985), who observed that: 'What is just . . . is what is proportional.'

The theory says that people decide what is fair using the principle that each person's gain over the *status quo* should be proportional to the appropriate social index for that person in the relevant context. The fair outcome generated by such an egalitarian norm will generally be very different from the outcome generated by a utilitarian norm. The latter is determined by dividing each player's gain by the appropriate social index. The sum of these corrected payoffs is then maximised. Aside from other significant differences, a player gets more from the egalitarian norm if his social index is increased, but less from the utilitarian norm.

14. Moral relativism

Long ago, Xenophanes made an empirical observation which says everything that needs to be said about the supposedly universal character of the various supernatural entities that have been invented down the ages:

> The gods of the Ethiopians are black and flat-nosed, and the gods of the Thracians are red-haired and blue-eyed.

[7] See, for example, Adams *et al.* (1963, 1965, 1976), Austin and Hatfield (1980), Austin and Walster (1974), Baron (1993), Cohen and Greenberg (1982), Furby (1986), Homans (1961), Mellers (1982), Mellers and Baron (1993), Messick and Cook (1983), Pritchard (1969), Wagstaff *et al.* (1992, 1994, 2001), Walster *et al.* (1973, 1975, 1978).

However, the fact that we all belong to the same species implies that some of our natural properties must be universal.

I think that one of these universal natural properties is the deep structure of fairness. If I am right, then all the fairness norms we use successfully in solving the small-scale coordination problems of everyday life are rooted in Rawls' original position. Space precludes giving the arguments, but a testable consequence is that we should expect all fairness norms that are actually used in all well-established societies to respond in the same way to changes in contextual parameters like need, ability, effort, and social status (Binmore 2005: chapter 11).

Although I believe that the deep structure of fairness is probably universal in the human species, the same cannot be true of the standard of interpersonal comparison that is needed to operate the device of the original position. This must be expected to vary, not only between cultures, but between different contexts in the same culture. Otherwise it would not be possible to explain the substantial differences in what is regarded as fair in different places and times, as documented in books like Elster's *Local Justice* (1992).

If I am right, the analogy with language is therefore close. All our fairness norms share the same deep structure, but just as the actual language spoken varies between cultures and contexts, so does the standard of interpersonal comparison that determines who gets precisely how much of a surplus that is divided fairly. For example, my theory suggests that it will always be regarded as fair for a person with high social status to get a smaller share than a less exalted individual, but the exact amount by which their shares differ will depend on the cultural idiosyncracies of the society in which they live.

15. Reform

My theory of fairness is an attempt at a descriptive theory; it seeks to explain how and why fairness norms evolved. Karl Marx might respond that it is all very well seeking to understand society, but the point is to change it, and I do not disagree. I hope very much that the scientific study of how societies really work will eventually make the world a better place for our children's children to live in, by clarifying what kind of reforms are compatible with human nature, and which are doomed to fail because they are not.

As an example, consider the pragmatic suggestion that we might seek to adapt the fairness norms that we use on a daily basis for settling small-scale coordinating problems to large-scale problems of social reform. This is one of the few things I have to say that traditional moralists find half-way acceptable. But they want to run with this idea without first thinking hard about the realities of the way that fairness norms are actually used in solving small-scale problems. In particular, they are unwilling to face up to the fact that fairness norms did not evolve as a substitute for the exercise of power, but as a means of coordinating on one of the many ways of balancing power.

This refusal to engage with reality becomes manifest when traditionalists start telling everybody how they 'ought' to make interpersonal comparisons when employing the device of the original position. But if I am right that the standards of interpersonal comparison we actually use as inputs when making small-scale fairness judgements are culturally determined, then these attitudes will necessarily reflect the underlying power structure of a society. One might wish, for whatever reason, that these attitudes were different. But the peddling of metaphysical arguments about what would be regarded as fair in some invented ideal world can only muddy the waters for practical reformers who actually have some hope of reaching peoples' hearts. Nobody is going to consent to a reform on fairness grounds if the resulting distribution of costs and benefits seems to them unfair according to established habit and custom, whatever may be preached from the pulpit.

It is true that facing up to such facts requires recognising that it is sometimes pointless or counter-productive to urge reforms for which a society is not ready. What would anyone have gained by urging the abolition of slavery in classical times, when even Aristotle thought that barbarians were natural slaves? What of the emancipation of women at a time when even the saintly Spinoza took time out to expound on their natural inferiority? Instead of tilting at such windmills, I think reformers need to make a hard-nosed assessment of the nature of the current social contract, and all the possible social contracts into which it might conceivably be transformed by pushing on whatever levers of power are currently available. Only when one has seriously thought through this feasibility question is there any point in asking what is optimal.

This pragmatic attitude mystifies traditional moralists, who pretend not to understand how a naturalist like myself can talk about optimality at all. How do I know what is best for society? What is my source of authority? Where are my equivalents of the burning bush and the tablets of stone?

The answer is that I have no source of moral authority at all—but I think everyone else is in precisely the same boat. I know perfectly well that my aspirations for what seems a better society are just accidents of my personal history and that of the culture in which I grew up. If my life had gone differently or if I had been brought up in another culture, I would have had different aspirations. But I nevertheless have the aspirations that I have—and so does everyone else.

The only difference between naturalists and traditionalists on this score is that naturalists do not try to force their aspirations on others by appealing to some invented source of absolute authority. We do not need a source of authority to wish that society were organised differently. If there are enough people with similar aspirations sufficiently close to the levers of power, we can get together and shift the social contract just because that is what we want to do—and for no other reason.

Note. I am grateful to the Economic and Social Research Council and to the Leverhulme Foundation for funding this work through the Centre for Economic Learning and Social Evolution at University College London.

Discussion

Robert Sugden, *School of Economics, University of East Anglia*

Ken Binmore has had the difficult task of summarising, in a one-hour lecture to a non-specialist audience, a theory of natural justice which he first presented in a two-volume treatise, comprising almost a thousand pages of often technical argument. He has succeeded admirably: his lecture is characteristically thought-provoking and pugnacious, while being beautifully clear. My task, of delivering a short critical commentary on his argument, is hardly easier.[8] Where should I start?

Reading this theory of natural justice, I have the sense that it is the work of two different Binmores. Each has something important to say about justice and fairness, but I am not convinced that their respective ideas can be combined into a single theory.

The first Binmore is Binmore the homespun philosopher, the scourge of Platonists and Kantians, the disciple of David Hume. His position is summed up, with typical bluntness, on the first page of his book, *Natural*

[8] The ideas I present in this discussion are developed more fully in my article-length review of Binmore's treatise (Sugden, 2001).

Justice. He says that orthodox moral philosophy asks how we ought to live, but that is the wrong question. The authority of philosophers who claim insight into this question is 'conjured from nowhere'; they 'have no more access to some noumenal world of moral absolutes than the boy who delivers our newspapers' (2005: 1). The only source of what we call moral intuitions is our experience of the rules that in fact govern our moral behaviour. So, he says, we need to study the processes of biological and social evolution by which morality in fact evolves.

Binmore has made a lot of intellectual enemies by taking this position, especially by presenting himself as the boy who sees that the emperor has no clothes. Modern moral philosophers, of the kind that Binmore criticises, have found many sophisticated ways of engaging in non-naturalistic discussion about morality while denying they are claiming to have access to absolute moral truths. To such philosophers, Binmore is a bull in a china shop, an ignorant economist talking about things he does not understand. So it is with some nervousness that I declare that I think that Binmore the homespun philosopher is right.

The second Binmore is the Binmore that most economists know best, the master of mathematical game theory. In its original form, game theory was an analysis of the strategies that ideally rational agents would follow when playing games against one another (as in what chess players call 'chess theory'). Two of the high priests of this approach to game theory are John Nash and John Harsanyi. Over half a century ago, Nash (1950) asked the question: What would be the result of bargaining between two ideally rational agents? He constructed a very general model of a bargaining problem, defined in terms of the characteristics which (he claimed) were the only ones relevant for rational agents, and produced a beautiful proof that ideally rational bargainers would settle on a particular solution to this problem. A few years later, Harsanyi (1953) asked the question: What principles of morality would be accepted by ideally rational agents? His approach was similar to John Rawls's (1972) later 'original position', but presented in a more rationalistic and mathematical way. Each person's moral judgements correspond with what she would prefer in a hypothetical position in which she did not know her own identity. This approach transforms moral philosophy into rational choice under uncertainty. The second Binmore is a disciple of Nash and Harsanyi.

The substantive content of Binmore's theory of natural justice is an intricate blend of the ideas of Nash and Harsanyi, with an added dash of Rawls. Moral principles are derived by imagining an original position in which the contracting parties do not know who they will become: each party is equally likely to become any one of the real people in their society. So far, Binmore follows Harsanyi. But where Harsanyi assumes that all the contracting parties make the same interpersonal comparisons of utility, Binmore allows each party to make his or her own comparisons, so they need to resolve their disagreements by negotiation. This is where Nash's theory of rational bargaining is put to use. Binmore also differs from his predecessors in allowing each of his imaginary contracting parties the right to demand a re-run of the whole exercise if, after finding out who she has become, she does not like the result. But then Binmore proposes the empirical hypothesis that this elaborate edifice of rational choice theory is also the conception of fairness that in fact has evolved in human societies. Thus, Binmore the mathematical game theorist and Binmore the homespun philosopher converge on the same moral theory. I suggest that points of greatest tension in Binmore's analysis come at the joins between these two approaches—where he tries to convince us that the forces of social evolution have selected the Harsanyi–Nash theory.

On Binmore's account, moral theory is selected to resolve coordination problems: in the evolutionary sense of the word, resolving these problems is the *function* of moral theory. One of his examples of the kind of problem he has in mind is: 'Who gives way to whom when cars are manoeuvring in heavy traffic?' So let us think more about this problem.

Suppose that you and I are drivers approaching a crossroads from opposite directions. You want to go straight ahead; I want to turn right across your path. Who gives way to whom? Binmore's answer seems to be that when I am deciding whether to give way to you, I imagine a hypothetical original position and reconstruct a rational bargain between you and me, taking account of my own judgements about interpersonal comparisons of utility between you and me, and of what I think your interpersonal comparisons are. If we succeed in coordinating, it is because our simulations of this hypothetical bargain lead to the same conclusion.

But can this really be how drivers decide whether to give way? In my experience of the crossroads problem, the driver who is first to reach the junction has priority; when there are queues on both roads, this results in the drivers on the two roads taking turns. I certainly follow this rule. Why? Because I have learned to expect other drivers to do so. Given this expectation, it is normally in my interest to follow the rule. Further, because I have come to expect the rule to be followed, and because I have good reason to believe that other drivers have the same expectation, attempts to seek advantage by deviating from the rule appear to me as unfair. In order to explain why I give way when I do, there is no need to consider hypothetical contracts; I am simply following the convention that has already evolved. In order to explain why I perceive breaches of this convention as unfair, there is no need to show that the convention would have been chosen in some idealised bargaining position; it is sufficient to recognise that the convention is generally followed, and that given that this is the case, each of us is harmed if the other breaches it.

That still leaves the question of how that convention has evolved. The answer, I suggest, is that it has evolved through the combined effects, over a long period, of people reasoning in just the way I do at the crossroads, each trying to do the best he can, given the behaviour of other drivers. Why the particular convention of priority to the first arrival? Probably because it's easy to learn, simple to use, applicable in a wide range of situations, and so capable of spreading from one situation to another by analogy. Such equilibrium selection mechanisms have very little to do with Nash bargaining theory, interpersonal comparisons of utility or original positions. They have very little to do with morality, as that has been understood by moral philosophers. But they can still generate rules that we come to regard as fair.[9]

My diagnosis is that Binmore is looking at morality from the perspective of a game theorist, imbued with the rationality-based traditions of that theory, rather than from the perspective of an empirical social scientist, trying to explain the facts on the ground. What he is trying to naturalise is not the collection of norms which we in fact find in human societies, but a form of morality which derives from an analysis of a world of ideally rational agents. Binmore the homespun philosopher is too much under the influence of the Binmore the mathematical game theorist. Dare I say that Binmore is too Kantian to be a true disciple of Hume? He has shaken off the myth of the truly good, but he is still in thrall to the myth of the truly rational. He has not accepted the full implications of a proposition to which his naturalism commits him: that even the greatest game theorists, even Nash and Harsanyi, have no more access to the concept of what is truly rational than the boy who delivers the newspapers.

[9] In my book, *The Economics of Rights, Cooperation and Welfare*, I argue that many conventions and norms have evolved in this kind of way (Sugden, 2004).

Ken Binmore, *Reply to Robert Sugden*

The first edition of Bob Sugden's *Economics of Rights, Cooperation and Welfare* (Sugden 2004) was one of the books that inspired my work on the evolution of fairness norms, and so it is comforting that his criticism of my approach seems to have softened since he last commented on my theory (Sugden 2001). I continue to believe that we are actually singing from the same hymn sheet, but there is space here only to consider two of the differences between us that he identifies in his commentary and elsewhere.

Sugden defends the Humean view that social norms are usually established gradually over time by a hard-to-model process of cultural evolution. I also believe this to be the case. Where we differ is in his belief is that there is nothing special about fairness norms. Brian Skyrms holds a similar view (Skyrms 1996, 2003). On this subject, their philosophy is more homespun than mine, because I believe that the Kantians John Rawls (Rawls 1972) and John Harsanyi (Harsanyi 1977) put their fingers on something real about how human fairness norms work when they independently formulated the idea of the original position.

As Sugden comments, I make intellectual enemies by stoutly denying Kant's metaphysics when it might be wiser to remain silent, but I hope he will find few takers for the idea that I come in two contradictory varieties: a homespun Binmore and a quasi-metaphysical Binmore who abandons evolution for neoclassical economics when it suits his purpose. What he fails to see is that there are not two Binmores, but two ways of interpreting Nash equilibria: a rational interpretation and an evolutionary interpretation. It is because one can sometimes pass back and forward between these two interpretations that I think game theory has proved so successful in both economics and biology.

A pair of strategies is a Nash equilibrum in a game if each is a best reply to the other. The rational reason for caring about Nash equilibria can be expressed in terms of what should be written in an authoritative book on how games should best be played. Such a book cannot recommend something other than a Nash equilibrium as the rational solution of a game, because at least one player would then have a reason for not following the book's advice. However, a book cannot be authoritative on what is rational if rational people do not play as it recommends.

The evolutionary reason for caring about Nash equilibria arises when the payoffs in a game correspond to how fit the players are. An adjustment process that favours the more fit at the expense of the less fit will stop working when we get to a Nash equilibrium, because all the survivors will then be as fit as it is possible to be in the circumstances. This is why the requirements for a Nash equilibrium are built into the formal definition of evolutionary stability used by biologists (Maynard Smith 1982).

I share Sugden's distaste for the arrogance of certain neoclassical economists. I agree that a methodology which exploits the dual interpretation of Nash equilibria deserves to be classed as speculative. But there is a baby in the bathwater that we cannot afford to throw away.

References

Adams, J. (1963), 'Towards an understanding of inequity', *Journal of Abnormal and Social Psychology*, 67: 422–36.

Adams, J. (1965), 'Inequity in social exchange'. In L. Berkowitz (ed.), *Advances in Experimental Social Science*, 2 (New York).

Adams, J. and Freedman, S. (1976), 'Equity theory revisited: Comments and annotated bibliography'. In L. Berkowitz (ed.), *Advances in Experimental Social Science*, 9 (New York).

Alcock, J. (2001), *The Triumph of Sociobiology* (New York).

Aristotle (1985), *Nicomachean Ethics*. Trans. T. Irwin (Indianapolis).

Aumann, R. and Maschler, M. (1995), *Repeated Games with Incomplete Information* (Cambridge, MA).

Austin, W. and Hatfield, E. (1980), 'Equity theory, power and social justice'. In G. Mikula (ed.), *Justice and Social Interaction* (New York).

Austin, W. and Walster, E. (1974), 'Reactions to confirmations and disconfirmations of expectancies of equity and inequity', *Journal of Personality and Social Psychology*, 30: 208–16.

Axelrod, R. (1984), *The Evolution of Cooperation* (New York).

Bailey, R. (1991), 'The behavioral ecology of Efe pygmy men in the Atari Forest, Zaire', Technical Report Anthropological Paper 86, University of Michigan Museum of Anthropology.

Baron, J. (1993), 'Heuristics and biases in equity judgments: A utilitarian approach'. In B. Mellors and J. Baron (eds.), *Psychological Perspectives on Justice: Theory and Applications* (Cambridge).

Binmore, K. (1994), *Playing Fair: Game Theory and the Social Contract, I* (Cambridge, MA).

Binmore, K. (1998), *Just Playing: Game Theory and the Social Contract, II* (Cambridge, MA).

Binmore, K. (2005), *Natural Justice* (New York).

Cohen, R. and Greenberg, J. (1982), 'The justice concept in social psychology'. In R. Cohen and J. Greenberg (eds.), *Equity and Justice in Social Behavior* (New York).

Damas, D. (1972), 'The Copper Eskimo'. In M. Bicchieri (ed.), *Hunters and Gatherers Today* (New York).

Dawkins, R. (1976), *The Selfish Gene* (Oxford).

Elster, J. (1992), *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens* (New York).

Erdal, D. and Whiten, A. (1996), 'Egalitarianism and Machiavellian intelligence in human evolution'. In P. Mellars and K. Gibson (eds.), *Modelling the Early Human Mind* (Oxford).

Evans-Pritchard, E. (1940), *The Nuer* (Oxford).

Furby, L. (1986), 'Psychology and justice'. In R. Cohen (ed.), *Justice: Views from the Social Sciences* (Cambridge, MA).

Fürer-Haimendorf, C. (1967), *Morals and Merit* (London).

Gardner, P. (1972), 'The Paliyans'. In M. Bicchieri (ed.), *Hunters and Gatherers Today* (New York).

Gauthier, D. (1986), *Morals by Agreement* (Oxford).

Hamilton, W. (1963), 'The evolution of altruistic behavior', *American Naturalist*, 97: 354–6.

Harsanyi, J. (1953), 'Cardinal utility in welfare economics and in the theory of risk-taking', *Journal of Political Economy*, 61: 434–5.

Harsanyi, J. (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (Cambridge).

Hawkes, K., O'Connell, J., and Burton-Jones, N. (1993), 'Hunting income patterns among the Hadza'. In A. Whitten and E. Widdowson (eds.), *Foraging Strategies and Natural Diet of Monkeys, Apes and Humans* (Oxford).

Helm, J. (1972), 'The Dogrib Indians'. In M. Bicchieri (ed.), *Hunters and Gatherers Today* (New York).

Homans, G. (1961), *Social Behavior: Its Elementary Forms* (New York).

Hume, D. (1975 [1777]) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd edn., ed. L. A. Selby-Bigge, rev. P. Nidditch (Oxford).

Hume, D. (1978 [1739]), *A Treatise of Human Nature*, 2nd edn., ed. L. A. Selby-Bigge, rev. P. Nidditch (Oxford).

Isaac, G. (1978), 'The food-sharing behavior of protohuman hominids', *Scientific American*, 238: 90–108.

Kaplan, H. and Hill, K. (1985), 'Food sharing among Ache foragers: Tests of explanatory hypotheses', *Current Anthropology*, 26: 223–45.

Knauft, B. (1991), 'Violence and sociality in human evolution', *Current Anthropology*, 32: 223–45.

Lee, R. (1979), *The !Kung San: Men, Women and Work in a Foraging Society* (Cambridge).

Lorenz, K. (1997), *King Solomon's Ring* (New York).

Mackie, J. (1977), *Ethics, Inventing Right and Wrong* (London).

Maynard Smith, J. (1982), *Evolution and the Theory of Games* (Cambridge).

Megarry, T. (1995), *Society in Prehistory: The Origins of Human Culture* (London).

Meggitt, M. (1962), *Desert People: A Study of the Walbiri Aborigines of Central Australia* (Chicago).

Mellers, B. (1982), 'Equity judgment: A revision of Aristotelian views', *Journal of Experimental Biology*, 111: 242–70.

Mellers, B. and Baron, J. (1993), *Psychological Perspectives on Justice: Theory and Applications* (Cambridge).

Messick, D. and Cook, K. (1983), *Equity Theory: Psychological and Sociological Perspectives* (New York).

Nash, J. (1950), 'The bargaining problem', *Econometrica*, 18: 155–62.

Pritchard, R. (1969), 'Equity theory; A review and critique', *Organizational Behavior and Human Performance*, 4: 176–211.

Rawls, J. (1972), *A Theory of Justice* (Oxford).

Riches, D. (1982), 'Hunting, herding and potlatching: Toward a sociological account of prestige', *Man*, 19: 234–51.

Rogers, E. (1972), 'The Mistassini Cree'. In M. Bicchieri (ed.), *Hunters and Gatherers Today* (New York).

Sahlins, M. (1974), *Stone Age Economics* (London).

Singer, P. (1980), *The Expanding Circle: Ethics and Sociobiology* (New York).

Skyrms, B. (1996), *Evolution of the Social Contract* (Cambridge).

Skyrms, B. (2003), *The Stag Hunt and the Evolution of the Social Structure* (Cambridge).

Sugden, R. (2001), 'Ken Binmore's evolutionary social theory', *Economic Journal*, 111: F213–48.

Sugden, R. (2004 [1986]), *The Economics of Rights, Cooperation and Welfare*, 2nd edn. (Basingstoke).

Tanaka, J. (1980), *The San Hunter-Gatherers of the Kalahari Desert: A Study of Ecological Anthropology* (Tokyo).

Trivers, R. (1971), 'The evolution of reciprocal altruism', *Quarterly Review of Biology*, 46: 35–56.

Turnbull, C. (1965), *Wayward Servants* (London).

Wagstaff, G. (1994), 'Equity, equality and need: Three principles of justice or one?', *Current Psychology: Research and Reviews*, 13: 138–52.

Wagstaff, G. (2001), *An Integrated Psychological and Philosophical Approach to Justice* (Lampeter).

Wagstaff, G., Huggins, J., and Perfect, T. (1996), 'Equal ratio equity, general linear equity and framing effects in judgments of allocation divisions', *European Journal of Social Psychology*, 26: 29–41.

Wagstaff, G. and Perfect, T. (1992), 'On the definition of perfect equity and the prediction of inequity', *British Journal of Social Psychology*, 31: 69–77.

Walster, E., Berscheid, E. and Walster, G. (1973), 'New directions in equity research', *Journal of Personality and Social Psychology*, 25: 151–76.

Walster, E. and Walster, G. (1975), 'Equity and social justice', *Journal of Social Issues*, 31: 21–43.

Walster, E., Walster, G., and Berscheid, E. (1978), *Equity: Theory and Research* (London).

Wilkinson, G. (1984), 'Reciprocal food-sharing in the vampire bat', *Nature*, 308: 181–4.

Wilson, J. (1993), *The Moral Sense* (New York).