politics, philosophy & economics article

London
Thousand Oaks, CA
and New Delhi

1470-594X
200602 5(1) 33–50

The evolution of fairness norms: an essay on Ken Binmore's *Natural Justice*

Paul Seabright
University of Toulouse, France

abstract This article sets out and comments on the arguments of Binmore's *Natural Justice*, and specifically on the empirical hypotheses that underpin his social contract view of the foundations of justice. It argues that Binmore's dependence on the hypothesis that individuals have purely self-regarding preferences forces him to claim that mutual monitoring of free-riding behavior was sufficiently reliable to enforce cooperation in hunter-gatherer societies, and that this makes it hard to explain why intuitions about justice could have evolved, since in such a society intuitions about justice would have had no adaptive advantage. I argue that it is empirically plausible that human beings display systematic other-regarding preferences (even if these are not always very strong). These could be incorporated into Binmore's general framework in a way that would enrich it and make it more useful for solving practical problems about justice.

keywords natural justice, fairness, norms, evolution, self-regarding preferences, Rawls, social contract

1. Introduction: the argument of the book

Ken Binmore's new book, *Natural Justice*,[1] is an ambitious attempt to explain why human beings have and use norms of distributive justice, as well as to justify for our use in modern societies a broadly egalitarian conception of such norms. It is a naturalistic enterprise, in two senses. First, it views norms of justice as having evolved as a part of human social behavior (rather than being implanted in us by God or as arising from some human capacity to perceive

DOI: 10.1177/1470594X06060618
Paul Seabright is Professor of Economics at the University of Toulouse, IDEI, 21 Allée de Brienne, F-31000 Toulouse, France [email: seabrigh@cict.fr]

timeless ethical truths). Second, it claims that we should understand the most important features of these norms with reference to the specific conditions of human evolution, rather than simply as an accidental by-product of that evolution. Our norms are the way they are because only in that way could they have survived given the conditions under which our species lived. One could imagine explanations of other aspects of human social life that were naturalistic in only the first sense. Our ability to enjoy music, for instance, might be an evolved capacity (rather than, say, a divine inspiration), without our necessarily being able to track the structure of our musical experience against the details of human beings' prehistoric behavior in the African savanna woodland.[2] In contrast, Binmore's enterprise of explaining justice is firmly naturalistic in both senses, and with this I am in wholehearted agreement. As will become evident, I disagree with some of his specific hypotheses, but his general explanatory framework commands my enthusiastic assent.

The book advances a specific hypothesis about norms of distributive justice, namely, that 'the rules of morality coordinate behavior among possible equilibria of social life'.[3] This claim has teeth. First, it says that defensible ethical reasoning recommends only actions that are consistent with self-enforcing (that is, equilibrium) configurations of behavior. For instance, morality does not tell us to engage in unilateral disarmament and rely on exhortation to keep the peace, except under some implausible scenarios about the likely consequences of doing so. Second, it says that, in most interesting social predicaments, many kinds of behavior *could* be self-enforcing, so we need a mechanism to coordinate on just one equilibrium out of the many. Third, it says that such coordination mechanisms typically have distributive consequences; they are not like the decision whether to drive on the left or the right, where either is as good as the other from everyone's point of view. They are more like the decision whether to register the car in your name or in mine, when this affects which of us gets to drive the car in the future – doing one or the other is better than not registering the car at all, but it makes a difference to both of us which one we do. Indeed, according to Binmore's theory, *all* the distributive conflicts over which the rules of morality have any purchase are in fact instances of choice between equilibria of our social life.

Underlying these explicit claims about what the rules of morality do are two hypotheses about the nature of the appropriate evolutionary explanation. First, Binmore believes that human beings have a universal genetic predisposition to be swayed by *some* set of moral rules, rules that guide action and that share a common deep structure of moral reasoning. He is almost certainly right about this, and it follows that such a predisposition must have evolved under natural selection because it proved adaptive for our hunter-gatherer ancestors. Only 10,000 years have passed since the invention of agriculture (too little time for our genetic endowment to have adapted significantly to post-hunter-gatherer life) and in any case, the biological evidence suggests that the last common maternal

ancestor of human beings alive today lived around 140,000 years ago, well before the end of the Paleolithic era. The mechanism of natural selection here is, incidentally, the Hamiltonian mechanism of kin selection, to which Binmore devotes a chapter. Second, the precise content of these rules is almost certainly *not* genetically determined, any more than the evidence for a genetic language instinct implies that we are genetically determined to speak French rather than Japanese. Rather, it undergoes cultural evolution, possibly through group selection, in the sense that norms are culturally transmitted from one generation of a society to another (via education, persuasion, and emulation), and the success of societies in competition with one another determines which norms spread most rapidly across the human population.[4]

Binmore advances not just claims about the kind of evolutionary explanation required, but also claims about the kind of moral rules that have in fact evolved under these evolutionary pressures. First, the common, deep structure of moral rules is, he claims, that which underlies the Rawlsian 'original position', namely, the thought experiment in which we abstract away from knowledge of our particular preferences and situation in life to consider that we might in some sense have been anyone. Second, Binmore claims that in hunter-gatherer societies the specific content of the moral rules seems to have been fairly strongly egalitarian (at least in terms of the distribution of material resources; he does not consider that such societies can nevertheless be fairly hierarchical in status terms). He also believes that broadly egalitarian rules are the only ones that can be defended in modern industrial societies. Why the post-hunter-gatherer, but pre-industrial societies that flourished for many millennia in between should have subscribed overwhelmingly to highly *inegalitarian* theories of justice is not, on this account, entirely clear.

Binmore shows (and this is the most powerful and original part of the book) that the device of the original position can be used in tandem with Nash bargaining theory to yield egalitarian substantive morality (at least egalitarian to the extent of the Rawlsian difference principle). In his view, our moral rules for resolving distributional conflicts are those that would emerge as the Nash solution to a bargain conducted between individuals behind a veil of ignorance as to which of the possible individuals in society they would turn out to be. There are two important qualifications to this claim. First, individuals must have empathetic preferences – when imagining themselves in the situation of others, they must imagine having the preferences of those others, not their own. Second, any solution must be incentive compatible once people know their true identities. In particular, individuals who turn out to have less favorable allocations of resources must not have an incentive to walk away from the social contract altogether. This is what explains why the original position yields a more egalitarian substantive theory than utilitarianism. It may also explain, incidentally, why societies in between the hunter-gatherers and our modern times were so inegalitarian, since many of them were slave economies. Slaves cannot walk

away from the social contract, unlike disgruntled hunter-gatherers or modern industrial workers who, for different reasons, cannot do their jobs well if encumbered with a ball and chain.

Overall, Binmore's theory represents a combination of the ambition of Rawls with the naturalism of Hume. He is less patient with the arguments of his opponents than Rawls, and his sarcasm is less deft than Hume's, but the combination of perspectives is remarkable, and deserves to be taken very seriously.

2. A tension in Binmore's argument

The two general hypotheses that Binmore advances about the evolution of morality (genetic evolution by kin selection for a general moral capacity and cultural evolution for the specific content of moral rules) seem plausible enough on the face of it, and indeed they are very probably true.[5] However, I shall argue that there is an important tension between them and his claim that morality serves to coordinate between the possible equilibria of social life. This tension arises because of the conditions that, according to Binmore, must hold for behavior to constitute an equilibrium of social life. I shall also argue that the tension can be resolved if we adopt alternative, more reasonable conditions for social equilibrium. The difference will matter a great deal, as we shall see.

What, then, are Binmore's conditions for behavior to constitute an equilibrium of social life? They are essentially the conditions for the folk theorem of repeated games: individuals must interact repeatedly, must not discount the future too heavily, must observe whether others are keeping to the equilibrium behavior, must have at their disposal means of retaliation that hurt the defector without being too costly to those inflicting them, and so on. Furthermore, and in contrast to the views of Herbert Gintis in this issue,[6] Binmore believes that individuals have essentially self-regarding preferences, and are not motivated except occasionally, erratically, and unreliably by such emotions as spite, or altruism toward strangers. Thus, behavior must constitute an equilibrium even though all individuals will seek actively to cheat at every available opportunity if they believe it will further their own interests.

Actually, the term 'self-regarding' is slightly misleading here – Binmore believes individuals are motivated to further what they perceive as their own individual interests *and* those of their kin, where the weight of their kin's interest approximates the Hamiltonian index of genetic relatedness (one-half for children and siblings, one-quarter for nephews and grandchildren, and so on). Rather than use the ungainly term 'self-and-kin-regarding' for such preferences, I shall use 'self-regarding' to describe preferences that include kin altruism, but exclude placing any intrinsic weight, positive or negative, on the welfare of non-kin. This is the key distinction, for the whole problem of explaining human sociality lies in understanding the roots of our elaborate cooperation with non-kin. Cooperation with kin is everywhere in the animal kingdom, but cooperation with un-

related members of the same species is, with minor exceptions, a uniquely human phenomenon.[7]

Thus, according to Binmore, when individuals engage in cooperative behavior with non-kin, they do so out of a fear of punishment by others if they were to behave otherwise, rather than from some intrinsically cooperative disposition. (It is important to note here that the punishment they fear may come not from those directly harmed by the individual in question, but rather from other members of the society who are engaging in 'equilibrium punishment', themselves motivated by fear of 'second-order punishment' if they fail.) Binmore appears to believe that these conditions (the folk theorem with self-regarding preferences) were broadly observed in hunter-gatherer societies when our moral capacities were evolving genetically, and, with some exceptions, are broadly observed in industrial societies today. For instance, he argues that the reason why children care for their elderly parents is that 'they would be censured by their community if they were to fail to carry out the role assigned to them by the social contract'.[8] This implies, of course, that the degree of mutual monitoring carried out by members of a society is extremely high.

The problem with such a view of cooperation is this. If mutual monitoring was as high as this implies, if individuals in hunter-gatherer societies were interacting frequently, observing each other's behavior closely, and threatening credible retaliation against free riders in ways that made cooperation always a social equilibrium even for people with entirely self-regarding preferences, then what was the point of a theory of justice? Why could it possibly have been adaptive to develop intuitions about just distribution? Why could not naked bargaining, based on relative bargaining strengths, do the job of coordinating on one of the many possible equilibria of social life? Nowhere in Binmore's book is it ever explicitly argued that a genetic predisposition to be swayed by moral rules would have proved adaptive in the probable conditions under which our hunter-gatherer ancestors lived. It is merely asserted that the 'deep structure [of social norms] is biologically determined, and hence universal in the human species'.[9] But no model, no evolutionary argument, is offered to explain why a gene determining our adherence to such a deep structure of norms would have survived and spread. Indeed, it seems probable that any such gene, had it emerged, would have led its ethical owner to fare badly against rivals who looked out more consistently for their own interests – who bargained using their own preferences rather than having anything to do with the veil of ignorance.

The only hint offered is a statement that 'fairness evolved for use in situations in which face-to-face bargaining isn't an option'.[10] However, if our hunter-gatherer ancestors were confronted sufficiently often with situations in which face-to-face bargaining was not an option, it is hard to believe that they can have been under sufficiently rigorous surveillance by others to enforce the rules of morality even against cheating by those who were continually on the lookout for opportunities to free ride. Herbert Gintis in this issue argues persuasively that it

takes only a relatively small degree of unreliability in the retaliation mechanism (due, say, to imperfect observability of the actions of others) for cooperation to break down dramatically. I am persuaded by this argument and will not repeat it here.

So it seems that Binmore is caught on the horns of a dilemma:

- Either mutual monitoring was more or less complete in hunter-gatherer societies, in which case a moral sensibility would have had no adaptive advantage, and some adaptive disadvantage, against a propensity for straightforward self-interested bargaining, or
- Mutual monitoring was significantly less than complete, so a moral sensibility had an advantage in coordinating on equilibria when members of the community could not bargain directly; but in this case, the equilibria between which they bargained would have had to be proof against cheating by very determined free riders who were often unobserved, a condition that would rule out many forms of cooperation as we know them.

There is a solution to this dilemma which I shall set out below. First, however, let me mention a theoretically possible, but (in my view) empirically implausible alternative solution. Conceivably, the degree of mutual monitoring that was necessary to enforce equilibria of social life among hunter-gatherers might have been significantly lower than that required for bargaining over which equilibrium to coordinate on. For instance, it might be enough, for enforcement of equilibrium, that each cheater be observed by one or at most a small subset of other members of the society, rather than by all other members. For bargaining over equilibria, however, all affected members might need to be involved, in order to establish clearly the rules by which everyone in the society is expected to play. Then it could be that hunter-gatherer societies had enough mutual monitoring to make cooperation an equilibrium, but not enough to make a moral sensibility redundant.

However, while conceivable, this strikes me as improbable, for two reasons. First, the degree of mutual monitoring required to enforce cooperative equilibria, given self-regarding preferences, is surely fairly high. It is not enough for cheating to be observable by one other member of the society, since that member has to be motivated to inflict punishment on the deviator by the fear of second-order punishment – so the observer must in turn be observed, and so on. Second, in order to coordinate on an equilibrium, it is hardly necessary for members of a society to be interacting frequently; they need only meet once in a while, in plenary session, so to speak. It is hard to imagine that societies that never met often enough to agree on the rules would, nevertheless, have had enough mutual monitoring for the rules to be enforceable.

3. A proposed solution

So what *is* the solution to the dilemma? First, it consists in accepting the hypothesis that the degree of mutual monitoring among hunter-gatherer societies (and a fortiori in modern industrial societies) fell a long way short of what would have been required to make cooperation an equilibrium of a game played by individuals with purely self-regarding preferences. It is true that, unlike in modern societies, most hunter-gatherers would have interacted most of the time with people they would expect to see again,[11] even when these people were not close kin. Nevertheless, many individuals would have found themselves often alone and unobserved, in circumstances in which there were significant private benefits to be gained from cheating. The circumstances of hunting and gathering (an uncertain environment, in which the location of game and plant food, and the manpower necessary to secure it, were constantly unpredictable) would have made that inevitable. Small groups would set out into the forest, would split up and rejoin each other; some individuals would head off in a variety of directions to cut off the escape of the prey, and so on. Someone who found an unexpected store of honey or berries could easily have eaten some and hidden the rest without telling the others. But if cheating was easy, suspicion and mistaken retaliation would almost certainly have been widespread. As an empirical fact about the organization of life in such societies, it seems to me almost inconceivable that the conditions of the folk theorem could have held if the preferences of individuals were entirely self-regarding.

So the second part of the solution consists in taking seriously the hypothesis that human beings evolved preferences that were not entirely self-regarding. For what it is worth, the experimental and other evidence that this is so strikes me as overwhelming, and Binmore's remarks on the subject are, by some margin, the least persuasive part of his book. Once again, I shall not repeat points made by Herbert Gintis, but add some further arguments that strengthen the case. These fall into two categories: first, arguments showing that some degree of other-regarding preferences could have dramatically altered the available equilibria of social life even when mutual monitoring was far from complete; and second, arguments that show how other-regarding preferences could have proved adaptive in hunter-gatherer societies and hence have been selected for genetically.

Let us take reciprocity as the main form of other-regarding preference in question (the question of whether it is the only relevant form need not detain us here). By 'reciprocity' I mean a tendency to respond to kindness with kindness and to cheating with revenge, *even when this is not what self-interest would recommend* (this is sometimes called 'strong reciprocity' to distinguish it from the kind of reciprocal behavior that merely tracks enlightened self-interest). Reciprocity is backward looking in that it responds to the *past* behavior of others and not, or not only, to their expected future behavior. To see what a difference the presence of other-regarding preferences could have made, note that it does not take a great

deal of capacity for reciprocity in a population for all members of the population to consider cooperating, even those who themselves are only self-regarding. First, by offering to share with someone else the food I have found, I may stimulate a reciprocity reaction that benefits me, whereas I would not do so if I knew that other people's behavior toward me in the future would be determined only by forward-looking considerations. Second, when some people are genuinely motivated by reciprocity, others may behave as though they were so motivated in order to gain a reputation for trustworthiness. It may take only a very small proportion of the genuinely motivated in the population to create reputational incentives for all or most of the others (this is a point well established in game theory, which makes it surprising that Binmore does not allude to it). Repeated game reasoning suggests, in other words, that even if formal structures of forward-looking incentives cannot entirely replace reciprocity, they can make a little reciprocity go a long way. Binmore may be dismissing reciprocity on the grounds that he cannot see how so little yeast can raise so much dough.

The benefits of reciprocity just described are community-wide, though, and accrue as much to the self-regarding as to the reciprocators themselves. Unless we appeal to group selection,[12] we still need an argument as to how reciprocity could have been individually adaptive. There is an obvious problem, which is that a tendency for reciprocity looks less adaptive than an alternative form of behavior that consists of responding to kindness with kindness and to cheating with revenge *only when this is what self-interest would recommend*. However, this objection is valid only on the assumption that other people have no way of observing whether someone else is a genuine reciprocator. Yet there would clearly be a significant adaptive advantage to a disposition for reciprocity which could signal its presence to others with some degree of reliability. For when it comes to inspiring the trust of others, the insensitivity of reciprocity to self-interested calculation is precisely its strength.[13] If I know that my present generosity to you will incline you to help me in the future, regardless of your interests at the time, I shall be more likely to take the risk of helping you. Your disposition for reciprocity makes you a more credible partner than someone who has no such disposition. This character, in short, gives you a power of commitment beyond the reach of the most sophisticated calculation. It is a power that the calculating can appreciate even though they cannot aspire to it – calculators would rather trust reciprocators than trust other calculators like themselves.

These considerations suggest that individuals who can simultaneously exercise trust shrewdly and inspire trust in others need to have some disposition for calculation in their dealings with others – but not too much, or no one will trust them. They also need some disposition for reciprocity – but not too much, or others will exploit them, and the memory of past wrongs will cast too long a shadow on their lives. They need a way to signal to others that they have the element of reciprocity that makes them trustworthy, and they need to signal it in a way that any purely calculating person could not convincingly mimic. The

psychologists Michael Owren and Jo-Anne Bachorowski have recently proposed an ingenious theory according to which smiling and laughter may have evolved in human beings as just such signals.[14] Both smiling and laughter are human capacities for which only the most rudimentary forms exist in other species. Both appear to signal emotions associated with a liking for others and a willingness to behave generously toward them – what psychologists call 'positive affect'. Both appear to *trigger* such feelings in others. Owren and Bachorowski suggest that the ability of smiling and laughter to act as reliable signals of positive affect, and therefore of trustworthiness, would have made them highly adaptive for the individuals that had these capabilities. Any genetic mutations favoring such capabilities would therefore have tended to spread. Given their reliability as signals of trustworthiness, evolution would also have tended to favor a tendency to respond warmly to them in turn.[15]

However, any signal that makes other people think I am trustworthy is one it would be highly useful for me to be able to fake. That way, I could make people trust me, and do favors for me, without incurring the cost of doing favors for them in return unless it suited me to do so. So, Owren and Bachorowski suggest, no sooner had smiling become reasonably well established as a reliable signal of trustworthiness than it also became adaptive to be able to make counterfeit smiles. Smiles that are under deliberate control are known to use a different set of neural circuitry than spontaneous smiles (the latter are also known as 'Duchenne' smiles after the 19th-century psychologist who first wrote about the difference). Not everyone can fake smiles successfully – indeed, politicians are predominantly drawn from among those human beings who can. But enough people can do so to suggest that the evolution of smile mimicry has proceeded quite far in the human species.

But almost nobody can fake laughter convincingly. Laboratory studies show that many people are unable to discriminate reliably between spontaneous smiles and those produced by good actors. Virtually everyone, though, can tell the difference between spontaneous laughter and the deliberate laughter of even very talented actors. Moreover, deliberate laughter provokes much less positive affect in those that hear it than does spontaneous laughter. These facts lead Owren and Bachorowski to suggest that laughter probably evolved later than smiling (no doubt in response to mimicry by the first prehistoric politicians). The fact that smiling was losing its reliability, because so many people could fake it, made it valuable to have another, better signal of positive affect. The possibility that laughter evolved later would explain why the ability to fake convincingly has not yet had time to evolve.

A telling piece of evidence in support of the signaling theory of laughter is the way in which, across all kinds of cultures in the world, people who have made a business deal with each other tend to seal the deal by having a drink together. Drinking alcohol notoriously affects people's judgment. In fact, alcohol is a depressant that not only makes people feel all stimuli less strongly, but particu-

larly diminishes people's sensitivity to pains, including future pains[16] (the reason why alcohol provokes car accidents is not primarily that it slows down people's reaction times, but much more significantly, that it makes them reckless, through diminished sensitivity to future dangers). In short, if people entering into a business relationship needed to keep a clear head in order to work out carefully how much they could afford to trust their new partners, having a drink together would be the worst possible way to seal a deal.[17] But alcohol is also, famously, a disinhibitor. Most importantly, it makes people laugh. Many business people in all cultures spend evenings exchanging jokes that, to begin with, virtually nobody finds funny, but at which everyone at the end of the evening is laughing uproariously. At the same time as it disables people's capacity for exercising trust wisely, alcohol enables people to inspire trust by stimulating that excellent signal of positive affect, namely laughter, that is not under direct voluntary control.[18]

Reciprocity is not the only mechanism that one can imagine playing the role of a visible commitment mechanism for trustworthy behavior. For instance, a religious sensibility might have evolved for similar reasons – a tendency to believe that spirits are observing your behavior even when no other human beings are doing so might well make individuals more trustworthy, and might also be hard for complete atheists to fake.[19] So a religious sensibility could be considered a genuine preference for cooperation over opportunism (albeit a preference underwritten by the promise of otherworldly rewards). The point is that the apparent adaptive advantage of opportunism over a genuine preference for cooperation may be offset by quite a small advantage of the genuine preference when it comes to signaling its own presence to others. Opportunists may do better than reciprocators when they get a chance to cheat, but unless they can hide the fact that they are opportunists, they may be given fewer opportunities to cheat in the first place. They may even face collective ostracism from the reciprocators within their group. All that is needed for such forms of other-regarding behavior to evolve is that there be sufficient correlation between the propensity for such behavior and the perception of it by others.

Finally, what does this account suggest about the origin of our theories of distributive justice? It is common ground between myself and Binmore that for societies to attain equilibria of cooperation requires them both to coordinate on an equilibrium and to enforce that equilibrium. Binmore suggests that hunter-gatherer societies would have had little difficulty in enforcing an equilibrium, but considerable difficulty in coordinating on one. A theory of distributive justice is for him the mechanism that evolved to bring about such coordination. On the alternative account I have outlined, the evolution of other-regarding preferences would have substantially enhanced the enforcement of cooperative equilibria. However, other-regarding preferences are not in themselves tantamount to theories of just distribution. So what were human beings doing developing such theories?

It seems quite likely that the broad contours of a theory of justice can be found simply by applying to an instinct for reciprocity the capacities for abstract reflection that modern human beings developed, probably around 150,000 years ago, and that have helped us to engage in symbolic manipulation of the kind that underlies mathematics, alphabetical writing, and is even present in some of the most striking cave art. After all, reciprocity is not just about symmetry between individuals (you are kind to me so I will be kind to you). In practice, it is also about symmetry between *circumstances* (you helped me when I was hungry, so I will help you when you are hungry). Reflecting systematically on symmetry between circumstances would very probably be adaptive in the sense that it would help human beings to calibrate their reciprocity reactions to the conditions under which they would be most effective. For instance, if you helped me when I was hungry, there is little point in my reciprocating in an undiscriminating fashion whenever I happen to have resources available. Much better to reserve my reciprocation to a time when it is of real value to you – when you are hungry too. This will require me to reason systematically and empathetically – to put myself in your position and imagine your preferences. Any systematic theory that enables me to do this more discriminatingly than on impulse is already an embryonic theory of justice. Notice, though, that these adaptive benefits of a theory of justice are still in terms of the enforceability of equilibria and not in terms of coordination between potential equilibria.

A theory of justice could also, of course, have been useful for the coordination purposes Binmore has suggested. Indeed, I would go further: there is quite a problem in understanding exactly how societies come to coordinate on equilibria of cooperation, in our own day as much as during our hunter-gatherer past. Binmore often writes as though there is a simple fact of the matter about what are the moral rules in a given society, thereby underestimating just how much contention and argument surrounds the rules – and has always surrounded them, in ancient and medieval times as well as in our own day. Globalization was already a fact of life several thousand years ago, and the first farming communities were having to fashion a basis for coexistence with others whose subscription to common standards could not be taken for granted. Under these circumstances, enforcing the rules could not wait for the establishment of a calm consensus as to what the rules were – on the contrary, those with political power or a capacity for rhetorical persuasion were engaged in arguing over the rules simultaneously with trying to enforce them. Such is the nature of the law – precedent, for instance, is created at the same time as specific legal decisions. Under these circumstances, too, anyone who advances a view as to what the rules should be will naturally be suspected of trying to distort the enforcement of the rules in their own favor. Those who can frame their arguments in the rhetoric of distributive justice, abstracting from the particular circumstances of the present case, may therefore enjoy a substantial persuasive advantage.

4. Can other-regarding preferences be reconciled with Binmore's theory?

Although Binmore is not always clear on this point, there are enough remarks scattered through the book, as through his two volumes on *Game Theory and the Social Contract*,[20] to suggest he would not dispute the empirical evidence that people frequently behave in ways that do not maximize their inclusive fitness, and that therefore we should not attribute to them exclusively self-regarding preferences. There are two main reasons why such evidence might nevertheless be considered irrelevant to Binmore's philosophical project, and both are vigorously argued by Don Ross in his article in this issue.

First, it can be argued that deviations from self-regarding preferences, while certainly frequent, are not sufficiently large, systematic, or reliable enough to be made the foundation of a theory of justice. Let me take these three features in turn.

1. *Deviations from self-regarding preferences are not (usually) large*. Binmore writes, and Ross cites: 'Apart from a few saints, who gives so much of their income to charity that it really hurts?'[21] I agree. However, as I argued in Section 3, they do not have to be large to make a large difference to the set of sustainable equilibria of the game of life, since in the presence of some behavior that is genuinely other-regarding, self-regarding preferences recommend much more prosocial behavior than in its complete absence. In short, though such deviations are certainly small, they are important.
2. *Deviations from self-regarding preferences are not systematic*. Binmore makes many dismissive remarks about behavioral economics (for instance, as needing 'a new utility function for each experiment'[22]). Here I strongly disagree. The evidence is strong that other-regarding preferences do not diverge randomly or capriciously from self-regarding preferences. Reciprocity, in short, has a structure. People cooperate in response to the generosity of others and punish when they observe behavior that strikes them as unfair. These divergences from self-regarding behavior are systematic enough to be predictable by other players in the game of life – and that is precisely why they can make large differences to the outcome of that game.
3. *Deviations from self-regarding preferences are not reliable*. Binmore makes a great deal of the evidence that repeated experience of playing one-shot prisoner's dilemmas tends to lead to more selfish play;[23] people converge, he claims, toward selfish play as they come to understand better the games they are playing. (While true of one-shot prisoner's dilemmas, however, this is less true of other games such as the ultimatum game.) This empirical conviction appears to be the main underpinning of Binmore's *methodological* adherence (which Ross emphasizes) to Hume's principle, namely, that human institutions should be proof against rational knaves.[24] Binmore and Ross both appear

> to interpret this principle as meaning 'Design human institutions to work in a world in which everyone is a rational knave.' However, Hume's principle is ambiguous. It could instead mean 'Design human institutions to be proof against invasion of the population by rational knaves.' I leave to one side the exegetical question of whether Hume himself considered this ambiguity to be important – I personally think it is crucial. As any evolutionary biologist would recognize, cooperation can be an evolutionary equilibrium if it is robust against invasion by noncooperative mutants, even if cooperative mutants could not in turn invade a noncooperative population. As a matter of fact and not of principle, I think most of our institutions of justice (police, the courts, and so on) are indeed reasonably robust against exploitation by individual sociopaths who care about absolutely nothing but their own inclusive fitness, but would break down entirely if everyone were like that. If every policeman behaved only according to what was in it for him and every judge decided every case according to what was in it for her, the social world as we know it would collapse, unless we succeeded in developing institutions of surveillance that were orders of magnitude more intrusive than those we actually have. It is precisely because a policeman behaving like that knows he risks coming before a judge who does not that the social architecture of modern life remains in place.

The second reason for considering deviations from self-regarding preferences as irrelevant to Binmore's project is that they might be regarded as short-run phenomena. In spite of the many pages that Binmore devotes to the distinction between the short, medium, and long run in his theory, I must confess to feeling that I still do not entirely understand it. The crucial question is whether deviations from self-regarding behavior in the short run serve to *enforce* equilibria in the game of life or merely to *coordinate* between alternative equilibria. On the account I gave above, we *must* understand them as serving, at least partly, an enforcement function, and this function must also be part of the explanation for why they evolved. Now Binmore may not be entirely consistent on this point, but he certainly often writes as though ethical norms serve only a coordinating function. For instance, in *Game Theory and the Social Contract*, Binmore writes that 'a player's duty simply lies in never deviating from the equilibrium path specified by the social contract', which might make one think that the norms of duty help to enforce that equilibrium. But he follows this immediately with the sentence: 'The custom of doing one's duty then survives because those who evade their obligation to honor the social contract suffer sufficient disapproval or punishment to make the deviation unattractive.'[25] This makes it clear that the role of duty is solely to indicate the equilibrium to be followed, the enforcement of which must be performed by a surveillance and punishment mechanism.[26]

As I have indicated, this is empirically implausible, but it is not central to Binmore's theory. Indeed, it may be that Binmore would be entirely happy to

grant that norms may also play an enforcement role, and to tone down those parts of his writings in which he suggests otherwise. The rest of his theory would be stronger, not weaker, for doing so.

5. Conclusion: What difference does it make?

How much does it matter which is right: Binmore's theory, as I have interpreted it in Section 1 or this alternative story? It matters a good deal, for both scientific and practical reasons. To take the scientific reasons first:

1. Taking seriously the presence of other-regarding preferences in human behavior gives us, as I have argued, a more empirically plausible account of how human cooperation with non-kin could have become established. It was surely some admixture of reciprocity with self-interest that enabled hunter-gatherer bands to take the first cautious steps toward conducting exchange with strangers. An itinerant trader making the first contact with an isolated band whose only motivation was self-interest would almost certainly have his goods stolen and would be lucky to be left with his life. After all, the band could hardly reason that offering him goods in return would be of benefit in the long run, since it would have no reason to expect him to come back again soon enough or often enough if it let him go, given that this was the first time the band had ever seen him. It has surely been reciprocity that, prehistorically, tipped the balance between hostility to strangers and a cautious willingness to deal with them. This was a truly momentous development in the history of humanity.
2. The fact that individuals have preferences that are not simply self-regarding makes much better sense of the way in which we talk about trust. If cooperation were due only to the folk theorem, we would not waste time trying to work out whether other people were trustworthy individuals. We would spend our time, instead, trying to see whether they had the right incentives to cooperate. We certainly do the latter, of course, but we also care a lot about the character of those with whom we interact. In sending a letter of reference for a potential employee, I am asked to write not just about ability, but also about honesty, reliability, commitment, and so on. If I wrote 'So-and-so is of very high ability, but his honesty will depend entirely on how well you monitor him,' even Ken Binmore might have doubts about employing the candidate in question.
3. The fact that reciprocity is backward looking can be reconciled with maximizing behavior by individuals on the assumption that the past behavior of others modifies my preferences toward them for the future. But it is not compatible with maximizing behavior under immutable preferences. Contrary to Binmore's dismissive remarks, cited above, about behavioral economics, this is not a weakness of the theory, but a strength. For it underlies the way in

which human social interaction is and has always been manipulative, in the ethically neutral sense that describes how adaptive behavior in human societies has depended upon shaping the behavior of others to an individual's own advantage. This has been a major factor in shaping our cognitive capacities, as Kim Sterelny has recently emphasized.[27] Human beings constantly manipulate each other's information, and it should not surprise us that, to the extent they are capable, they should seek to manipulate each other's *preferences* as well. This does not make our models of human preferences empty or epicyclic, any more than the presence of signaling and mimicry make our models of human beliefs empty or epicyclic.

The practical reasons why it matters whether Binmore's theory or this alternative account are correct have to do with our response to failures of cooperation, and with disputes about what to do. Taking these in turn:

1. Folk-theorem reasoning always prescribes increased monitoring or more severe punishment in response to cooperation failures, while in the presence of other-regarding preferences, such responses may be counterproductive. For instance, there is some evidence that increasing penalties for tax evasion may sometimes result in reduced tax compliance, because it is interpreted by previously honest taxpayers as a signal that many others are dishonest, thereby prompting them in a spirit of reciprocity to reduce their own compliance.[28] This is not to say that penalties do not matter, but cultures of compliance depend on more than monitoring and penalties. More generally, the design of effective systems of incentives requires us to treat other-regarding preferences, if they exist, not as purely orthogonal to the rest of incentive theory, but as requiring explicit integration into mechanism design.[29]
2. When we disagree with others about questions of justice, the route to a solution rarely consists simply of inviting our interlocutors to undertake a sociological inquiry to discover what are the rules of the society in which we live. Such an inquiry may be helpful, of course, but the rules are often ambiguous and in the process of evolution toward an uncertain future. For instance, if an African woman from a traditional rural community defies her mother's wish that she circumcise her daughter, is she simply making a mistake about what her society prescribes or is she arguing that her society and its rules are not, after all, those that her mother takes them to be, and that she conceives herself and her daughter as citizens of a wider and different society? Disputes about distributive justice may often be less emotive than this, but we will handle them more sensitively, I suspect, if we take seriously the way in which coordination upon equilibria, and enforcement of equilibria, are inextricably mixed up together.

In summary, therefore, I have a significantly different view from Ken Binmore (if I have interpreted him correctly) about some of the empirical circumstances in

which the moral capacities of human beings evolved, and in which they continue to be exercised today. This leads me to express strong doubts about his substantive theory. But our disagreement is already focused upon issues of fact, albeit ones that are hard to settle historically in a definite way; they do not rely on rival appeals to moral intuition or to outlandish thought experiments. For this reorientation of our moral arguments, Ken Binmore deserves the gratitude of all those who want to understand human societies and their tussle with ethical ideas.

notes

I am grateful to Jonathan Riley, Don Ross, and John Weymark for comments on an earlier draft, while absolving them from complicity in any errors of fact, logic, or interpretation that remain.

1. Ken Binmore, *Natural Justice* (Oxford: Oxford University Press, 2005).
2. For more on the evolutionary origins of music, see Norman M. Weinberger, 'Music and the Brain', *Scientific American*, November (2004): 66–73, and the references cited therein.
3. Binmore, *Natural Justice*, p.14.
4. It is possible also that some norms may spread easily in a population because they have certain properties (salience or ease of memorization) that are conducive to transmission without necessarily contributing to the fitness of either individuals or groups that adopt them. This is particularly likely to be true of those norms that tend to spread in today's information-rich societies, helping themselves to features of our psychology that evolved in the comparatively information-starved environment of the African woodland savanna. It also underlines that cultural evolution need not be driven by either individual or group selection: the norms of one group can be copied by a second group, even if the first group does not survive for long thereafter. Indeed, the invention of cultural artifacts, including writing, means that we can copy some of the norms of groups that no longer exist at all, as in neoclassical artistic movements and certain reactionary or new age philosophies. Perhaps (this is more speculative), the norms of vanished groups exert a particular fascination for us, since we no longer have to confront evidence about the problems associated with their implementation and can fantasize freely about their attractions.
5. Subject, that is, to the reservations about the group-selectionist version of cultural evolution expressed in Note 4. While Binmore seems broadly to endorse group selection for norms, he makes a number of remarks that suggest he would not dispute that norms can sometimes propagate themselves without contributing to group fitness. In any case, this issue can be considered separately from the other elements in his theory.
6. Herbert Gintis, 'Behavioral Ethics Meets Natural Justice', *Politics, Philosophy and Economics* 5(1): [PAGE NOS], in this issue.
7. I have explored this theme in much greater detail in Paul Seabright, *The Company of Strangers* (Princeton, NJ: Princeton University Press, 2004). Some of the arguments below draw on passages in that book.
8. Binmore, *Natural Justice*, p. 87.
9. Ibid., p. 18.

10. Ibid., p. 27.
11. The evidence suggests that prior to the adoption of agriculture interaction with strangers was rare, and where it took place, was far more often violent than peaceful, let alone cooperative. Seabright, *The Company of Strangers*, pp. 48–53 especially.
12. Some have done just this: see David Wilson and Elliott Sober, 'Re-Introducing Group Selection to the Human Behavioral Sciences', *Behavioral and Brain Sciences* 17 (1994): 585–654; Herbert Gintis, 'Strong Reciprocity and Human Sociality', *Journal of Theoretical Biology* 213 (2000): 103–19.
13. This argument is due originally to Robert Frank, *Passions Within Reason: The Strategic Role of the Emotions* (New York: Norton, 1988).
14. Michael Owren and Jo-Anne Bachorowski, 'The Evolution of Emotional Expression: A "Selfish-Gene" Account of Smiling and Laughter in Early Hominids and Humans', in *Emotions*, edited by Tracy Mayne and George Bonnano (New York: The Guilford Press, 2001), Ch. 5.
15. There are other theories of the evolution of laughter, not necessarily incompatible with the one outlined here. For instance, V.S. Ramachandran and Sandra Blakeslee, *Phantoms in the Brain* (New York: Harper Collins, 1999) propose that laughter evolved to signal to other members of a social group that a feared threat (from a predator, for instance) is in fact not serious; this could explain why we laugh in relief. This could in turn account for the use of laughter to signal to a listener that the person laughing is not himself a threat.
16. J.A. Gray and Neil McNaughton, *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System*, 2nd edn., Oxford Psychology Series No. 33 (Oxford: Oxford University Press, 2000), Ch. 4.
17. It is true that drinks tend to be served after agreements are signed, but trust is as important after the agreement as before: each party still needs to decide whether it can trust the other to stick to the agreement.
18. Mark Greenberg (personal communication) points out that what biologists call the 'handicap principle' may also be at work: by drinking alcohol in your company I am signaling that I am so confident in my skill at discerning your trustworthiness that I am willing to disable it with a powerful depressant drug. This can work both to reassure you that I intend to trust you (by behaving in a trustworthy manner myself) *and* to warn you how quickly you will be discovered if you betray that trust.
19. This might also explain why some religious organizations stress the high costs of membership – they make it harder to mimic religious commitment and therefore make the signal of character embodied in membership a more credible one.
20. Ken Binmore, *Game Theory and the Social Contract, Volume 1: Playing Fair* (Cambridge, MA: MIT Press, 1994); Ken Binmore, *Game Theory and the Social Contract, Volume 2: Just Playing* (Cambridge, MA: MIT Press, 1998).
21. Binmore, *Natural Justice*, p. 9.
22. Ibid., p. 60.
23. Ibid., p. 186, for example.
24. 'What theory of morals can ever serve any useful purpose unless it can show . . . that all the duties which it recommends, are also the true interest of each individual?' See David Hume, *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd edn. (Oxford: Clarendon Press, 1975),

p.280. Cited by Binmore, *Game Theory and the Social Contract, Volume 1: Playing Fair*, p. 30, fn. 28.

25. Binmore, *Game Theory and the Social Contract, Volume 2: Just Playing*, p. 277.
26. Ross also writes that 'People are modeled as coordinating their behavior in repeated-game equilibria around these norms.' See Don Ross, 'Evolutionary Game Theory and the Normative Theory of Institutional Design: Binmore and Behavioral Economics', *Politics, Philosophy and Economics* 1 (2006): [PAGE NOS], in this issue.
27. Kim Sterelny, *Thought in a Hostile World: The Evolution of Human Cognition* (Oxford: Blackwell, 2003).
28. See Dan Kahan, 'The Logic of Reciprocity: Trust, Collective Action and Law', Working Paper No. 281 (New Haven, CT: John M. Olin Centre for Studies in Public Policy, Yale Law School, 2002).
29. I have developed economic models of incentive mechanisms involving other-regarding preferences in two recent papers: Paul Seabright, 'Continuous Preferences can Cause Discontinuous Choices: An Application to the Impact of Incentives on Altruism', Discussion Paper No. 4322 (London: Centre for Economic Policy Research, 2004); Paul Seabright and Colin Rowat, 'Intermediation by Aid Agencies', *Journal of Development Economics* (forthcoming).