Analyse & Kritik 28/2006 (© Lucius & Lucius, Stuttgart) p. 32–56

Jonathan Riley

Genes, Memes and Justice

Abstract: Ken Binmore argues that justice consists in a proportional bargaining equilibrium of a 'game of morals', which corresponds to a Nash bargaining equilibrium of a 'game of life'. His argument seems unassailable if rational agents are predominantly self-interested, an assumption that he is apparently willing to make on the grounds that human behaviour is ultimately constrained in accord with the selfish gene paradigm. But there is no compelling scientific evidence for that paradigm. Rather, human nature appears to be highly plastic. If so, rational agents might eventually be moulded by cultural forces into social and moral actors who effectively believe that they are the same person—no different from anyone else—when it comes to certain vital personal interests which ought to be treated as rights. In this context, a utilitarian outcome is an efficient and fair equilibrium of the game of life. Compliance with the rules is enforced by the actor's own conscience, a powerful internal 'judicious spectator' which threatens to inflict harsh punishment in the form of intense feelings of guilt for cheating.

0. Introduction

Ken Binmore, drawing inspiration chiefly from David Hume but also to some extent from John Rawls and John Harsanyi, proposes what he calls a "modern whig" moral and political theory according to which "the way to a better society lies in appealing to the rational self-interest of all concerned" (186). New whigs of Binmore's stripe argue that rational agents can be expected to reconcile fairness with their self-interest, at least temporarily, because certain 'common understandings' of what's fair are accepted as a result of biological and cultural evolution. True, the common norms of fairness or justice will reflect to some extent the unequal bargaining power of the agents. Even so, the norms demand that any agent should receive a fair share of any unanticipated social surplus, or bear a fair share of any unanticipated social deficit, where 'fair' means that his share is proportional to his 'worthiness' as conventionally assessed in light of his bargaining position. The ratio of any agent's fair share to his worthiness is equal to that of any other agent. By prudently building the common understandings of fairness into the design of social institutions, including duly regulated markets as well as political constitutions, Binmore insists, enlightened whig planners could achieve fair distributions of scarce resources by mutual consent, without sacrificing Pareto-efficiency.

At the same time, neo-whigs of Binmore's persuasion maintain that rational agents cannot be relied upon to be fair for very long, if there are no unanticipated changes in the feasible set of outcomes. If such changes don't occur, these agents

will lose interest in being fair to one another: all traces of morality will be leached out of their norms of behaviour, and they will make use of their wealth, political connections, social status, and any other bargaining advantages to promote their selfish concerns as far as they can, perhaps even to the point of destroying mutual cooperation altogether, at least in their dealings with strangers as opposed to kin. Enlightened whig planners must also take account of this tendency of rational agents to lapse into knavery. Social institutions must be designed which rational self-interested agents have incentives to operate. Such agents must be deterred from deviating from the given institutional rules by credible threats of punishment from their fellows. The institutions must be self-enforcing in the sense that no external enforcement agency—god, for example, or some other omnipotent and omniscient 'impartial observer'—is needed to force agents to comply with the rules.

Binmore employs an evolutionary game-theoretic framework to present his neo-whiggery. I shall now discuss the main elements of his framework as I understand it, with a view to offering some objections and suggested revisions to his naturalistic theory of justice.

1. The Game of Life

Binmore models the 'game of life' as an indefinitely repeated game of complete information, in which rational expected utility maximizers value the future (that is, they discount future utility payoffs at suitably low rates) and thus can coordinate on any of multiple Pareto-efficient Nash equilibrium points. The central issue, as he sees it, is the selection of an efficient Nash equilibrium which is also a fair outcome.[1]

The agents' strategy choices are required to constitute a Nash equilibrium point because otherwise their choices will not be stable: some will have incentives to alter their strategies given the strategies of the others. The agents are also assumed to coordinate on an *efficient* Nash equilibrium because otherwise some agents' utility payoffs can be increased without making anyone else worse off. Given his suitably low discount rate, even a selfish agent recognizes that he and everyone else will do best if every individual chooses a strategy such as 'TIT-FOR-TAT' in which the individual reciprocates good for good as well as evil for evil.[2] True, it is not irrational to refuse to cooperate, and inefficient outcomes do remain among the equilibria of the indefinitely repeated game of life. Yet

[1] It is important to keep in mind that Binmore makes certain simplifying assumptions for ease of analysis. He ignores problems of incomplete information, for instance, and doesn't allow for the formation of coalitions. He even admits that it may become necessary to abandon the repeated game framework altogether "to consider stochastic games in which the strategies chosen by today's players can influence not only their own payoffs, but also the game played by tomorrow's players". Such stochastic games are "much harder to analyze", however, and it is likely that no "recognizable analogue of the folk theorem holds" for them if today's players can "restrict the options available in future games sufficiently quickly" (199).

[2] Binmore points out that the attention given to TIT-FOR-TAT is unfortunate because it suggests that reciprocity cannot work unless a player directly retaliates against another who fails to cooperate with him. But "the injured party" doesn't need to be the agent "who

any rational agent will cooperate if he has reason to believe that he or anyone else will be detected and suitably punished for failing to cooperate. If everyone believes that everyone is deterred from non-cooperation by a threat of sufficient punishment, everyone can be expected to cooperate.[3]

Even if they are willing to cooperate, rational agents can achieve any of multiple efficient equilibria. As Binmore conceives it, morality or justice is a device for selecting a fair outcome among the multiple efficient equilibria. This equilibrium selection device is a product in part of biological evolution, he suspects, and in part of cultural evolution. Although more needs to be said about it, the key point for the moment is that the device is an evolutionary tool that enables rational agents to quickly and smoothly come to a mutual agreement concerning the distribution of any social surplus or deficit that arises as conditions change in unanticipated ways. The feasible set of outcomes in the game of life can expand or contract repeatedly as a result of exogenous shocks such as technological innovations or natural disasters. The tool of morality is valuable for all concerned since it obviates the need for face-to-face bargaining while preventing internal dissent and conflict over distributive issues. By making use of this tool, rational agents can move immediately by mutual consent to an outcome which all of them regard as fair.

The game of life as depicted so far is highly flexible and open-ended because it is compatible with any substantive theory of human nature. Binmore suggests as much when he says: "Game theory remains the same whatever view one takes on the nature of human nature" (65). In particular, there is not yet any reason to conclude that rational agents are selfish, or that they will only agree to coordinate on an outcome that satisfies one criterion of fairness rather than another. Rather, up to now, a rational agent is assumed to behave in a logically consistent manner and, if all have sufficient reason to cooperate, to cooperate with others in a way that all agree is fair in some sense. Otherwise, preferences can be of any content, and there are no restrictions on the motivations underlying the preferences. A rational agent's behaviour simply reveals a complete and transitive preference ranking such that his preference for one possible outcome rather than another implies that the one outcome has more utility value than the other does for him, with his relative preference intensities reflecting how much more utility value the one outcome has in comparison to the other. Utility is merely a cardinal representation of his revealed preference ranking, whatever the content of the preference.

---

punishes a cheater in multiperson interactions" (79). Reciprocity can account for cooperation if cheaters are punished by third parties. For an illustration, see 86–89.

[3] The assumption of complete information is important here. It implies that an efficient equilibrium can be sustained at little cost because anyone who fails to cooperate is observed by the other players. The costs of detecting and punishing cheaters are low because everyone knows who they are. As Binmore says, this assumption may be acceptable "in the case of small bands of hunter-gatherers" but it isn't in the case of a big modern society. Nevertheless, he maintains the assumption, and thereby ignores inefficiencies associated with prohibitive costs of monitoring and punishing cheaters in large anonymous societies, because his "focus is on how fairness norms evolved in the first place" (82).

Nevertheless, to provide a moral and political theory that purports to explain how rational agents *ought* to behave if they are to achieve fair outcomes, something must be said about 'the nature of human nature' and about what people believe they are doing when they are cooperating fairly with one another. In other words, something must be said about the content of the preferences revealed by their consistent behaviour, "whether they are intending to [behave that way] or not", and thus about the interpretation of the utilities they are maximizing. At times, Binmore leaves the impression that he views the interpretation of utility payoffs in any context as merely an 'empirical question', with the implication that what the members of any society regard as desirable or fair behaviour can have any content. This impression is reinforced by his suggestion that a society's practices cannot rationally be criticized by outsiders as objectively bad: "Someone looking in from the outside or back from the future may not agree, but who are they to insist that their standards should take precedence over those of the society they are criticizing?" (173) Yet, all things considered, he does not really seem to have in mind such a merely empirical approach to the content of utilities.

Rather, Binmore apparently takes for granted that rational agents in every social context are at bottom selfish and inclined to take advantage of others, especially strangers, in the absence of any threat of suitable punishment by other people. As a result, rational agents will not accept any outcome as fair unless it is generated by a bargain which selfish agents with unequal bargaining power have incentives to enforce. Fairness takes the form of a self-enforcing bargain among selfish agents, and this in turn leads to the conclusion that a fair outcome is necessarily an egalitarian or proportional bargaining equilibrium.

Yet why does it make sense to assume that rational agents everywhere are ultimately selfish, thereby implying that they will never implement an outcome as fair unless it is a proportional bargaining equilibrium? Binmore apparently associates the axiom of selfishness with biological and cultural evolution. In particular, he suggests that the "deep structure" of the game of life is such that rational humans behave as if they are being instructed by their genes to maximize their reproductive fitness. He goes so far as to admit that his main motivation for even assuming that people "always choose consistently" is "the consideration that genes or memes that consistently promote their own replication are more likely to survive than those that don't" (118).

More generally, Binmore's view seems to be that human nature is properly understood in terms of economic, cultural and genetic processes operating on different time scales which he calls the short run, the medium run, and the long run, respectively:

> "Economic, social, and biological processes actually proceed simultaneously, but models which reflect this reality would be prohibitively difficult to handle. One therefore attempts to approximate the way the world actually works by assuming that short-run processes are infinitely faster than medium-run processes, and medium-run processes are, in turn, infinitely faster than long-run processes." (157)

As a result, the game of life exhibits a degree of complexity which requires further clarification. If I understand him correctly, Binmore is defending a pluralistic theory of human nature according to which rational agents are ultimately selfish yet the preferences they reveal can vary widely in content across different contexts.

2. The Long Run

Consider the long run, also known as 'biological time'. The idea seems to be that a human being, when viewed from the perspective of this time scale, is just a location for genes which are competing for survival with other genes in the indefinitely repeated game of life. Using terminology made popular by Dawkins (1976; 1982), 'selfish genes' compete over a period of generations to establish themselves in a host human population that is maximally adapted to reproduce the genes and enable them to survive in the given natural environment, ignoring changes to that environment for ease of analysis. Genes may be seen as biological algorithms for playing strategies that maximize the genes' replication. A rational human behaves as if he were instructed by his genes to consistently have more children, subject to any cultural and economic constraints that may happen to exist. Rational humans can thus be viewed as selfish agents playing strategies that maximize expected utility measured in terms of reproductive fitness.

It takes many generations to converge on a long-run equilibrium in accord with some dynamic process of natural selection. Suppose that the proportions of the human population playing distinct strategies in the next generation are related to the proportions of the population playing the strategies in the current generation in accord with the "replicator dynamics" (Taylor/Jonker 1978; Zeeman 1980). A strategy then gradually invades the population if the average reproductive fitness (utility payoff) of the individuals who play it in their interactions with the other players exceeds the average fitness of the population as a whole, that is, the average fitness of all the individuals who play any of the strategies in their interactions with others. New mutant strategies might be played in any generation, and then attempt to invade the population. The dynamic process will stop only if it reaches a Nash equilibrium. Roughly, it stops once an evolutionary stable strategy (possibly a mixed strategy) is established in the population.

Rational humans in 'biological time' behave, then, as if they are selfish genes seeking to replicate themselves in competition against other genes. Consistently with this, rational agents will cooperate with kin in accord with Hamilton's (1964) rule.[4] But these rational agents have no reason to cooperate with genetic 'strangers' unless threatened with sufficiently harsh punishment to endan-

[4] Given that there is a probability of 0.5 that brothers or sisters will pass on each other's genes to their children, for instance, and a probability of 0.125 that cousins will do so, selfish genes will dictate strategy choices that reflect these probabilities. In short, biological individuals seeking to replicate their genes will behave as if they augment their personal expected utilities (biological fitness) by adding the utilities of their relatives, weighted by the relevant probabilities. See 102–105, 108–112.

ger their own reproduction and survival.[5] If they do expect to receive such punishment for non-cooperation, then an efficient long-run equilibrium can be sustained. Such a long-run equilibrium may be seen as a Nash bargaining solution in which rational humans are behaving as if they are selfish genes employing whatever bargaining advantages they have to reproduce and survive, and there are no opportunities to improve the average reproductive fitness of the human population without making at least some individual worse off in terms of fitness.

It might be objected that individuals will find it too costly to carry out such harsh punishment, including physical injury and confinement, for non-cooperation with genetic strangers. But a fairly small band of individuals, including genetic strangers, may do so to sustain an efficient equilibrium among its members in a competition with other tribes for survival over the same territory. Given that they can monitor each other at little cost, perhaps as many as two hundred people would have reason to mutually cooperate, especially since the kin among them are more inclined to cooperate in any case. Eventually, perhaps, a population consisting almost entirely of kin could emerge as a result of one tribe producing more children than the others, one tribe annihilating the others, and/or intermarriage among the tribes.[6]

None of this is to deny that cultural forces also influence human behaviour, a point that Binmore often stresses. But he thinks it "obvious" that "human biology must impose constraints on what social contracts can evolve", and that "*universals*" in human behaviour can be found only if "we ... look beneath the differing cultures of different societies ... at the deep structure of human social contracts written into our genes" (13–14, emphasis original; see, also, 45–46). In his view, moreover, culture tends to erode away in the long run, leaving biological imperatives to hold sway.

The 'deep structure' of the game of life is apparently such that rational agents are universally driven by their genes to pursue the selfish goal of replicating the genes, whether or not the agents are conscious of this goal. Individuals will need to learn that this is the case, it seems, because, as Binmore admits, genes certainly do not directly program human behaviour in very many situations. Nevertheless, the genes do somehow determine their human host's utility payoffs when interacting with others, he claims. Rational humans act as if they calculate their utilities in accord with Hamilton's rule when interacting with kin, whereas

---

[5] In response to the objection that human beings share virtually all of their DNA, Binmore (103) argues that "we are never concerned with genes that are *always* present in the human body, but only with a particular piece of behavior that will be modified or left alone according to whether a recently mutated gene is present or absent. In line with the selfish gene paradigm, it is from the point of view of such a mutant gene that fitness must be evaluated—not that of the individual host in whose body the gene is carried."

[6] See 12, 67–68, 134–35. In connection with Rousseau's story of a stag hunt, for instance, he argues that there isn't much mystery why a band of hunters will cooperate with each other in a competition with other groups, "provided that one remembers that the group selection fallacy doesn't apply when each group is operating an *equilibrium* in the game its members play with each other". He points out that "societies in which men hunt cooperatively are more successful than societies in which they don't, because they produce more food overall". Moreover, "cooperative hunting can be sustained as an equilibrium by punishing men who don't pull their weight" (135). See, also, Skyrms 2003.

they do not do so when interacting with biological strangers except perhaps in the case of a few friends whom they largely confuse with kin. Rational humans will presumably come to appreciate these fundamental constraints imposed by their genes on their behaviour.

But biological processes operate very slowly relative to cultural and economic processes in Binmore's framework, and it remains an open question how much weight the genes are supposed to have relative to other forces in determining human behaviour in the medium run or short run.

3. The Medium Run

In the medium run, also known as 'social time', 'biological time' is assumed to be frozen and genes are regarded as parameters. A rational person's choices, insofar as they are motivated by his genes seeking to replicate themselves, are assumed to be fixed. Viewed from the perspective of 'social time', the person is apparently just a location for what Dawkins (1982) calls 'memes' which are competing for survival with other memes in the indefinitely repeated game of life. Memes are cultural analogues of genes, that is, bits of culture or cultural algorithms which can vary widely in content, including linguistic conventions, fashions of dress, standards of interpersonal comparisons, and so forth. Different memes compete over a period of months or years to establish themselves in a host human population that is maximally adapted to replicate the memes and enable them to survive in the given social environment. Rational humans can be viewed as agents playing strategies that maximize expected utility measured in terms of cultural fitness or social status.

It may take anywhere from a few months to many years to arrive at a medium-run equilibrium in accord with some dynamic process of cultural selection such as the replicator dynamics. But otherwise the mechanics of the process are similar to those of the biological process. Although they are not inherited, memes are assumed to spread from individual to individual through a dynamic process of imitation, which may be largely unconscious. Individuals perceived as successful, whether at getting wealth, power, fame, their needs satisfied, or some other valuable resource depending on the context, are imitated by others in their customs, habits and practices. The strategy choices dictated by the memes lodged in these successful individuals thereby tend to invade the population.

Rational humans in 'social time' or 'cultural time' behave, then, as if they are memes seeking to replicate themselves in competition against other memes. Without further argument, however, ambiguity remains as to why the strategy choices dictated by memes to replicate themselves should have a selfish cast. Since it would seem that memes could potentially be of any content, why should it be assumed that the memes which tend to spread among rational humans will be skewed in favour of the selfish concerns of successful individuals? An important point to keep in mind, however, is that these cultural norms are tending to wash away in the long run, except to the extent that they are compatible with the selfish biological imperatives that make up the 'deep structure' of the game

of life. Moreover, the influence of the genes on behaviour apparently remains sufficiently powerful in the medium run that an efficient medium-run equilibrium can still be seen as a Nash bargaining equilibrium. Despite his cultural veneer, it seems, the rational human remains at his core a selfish gene, seeking to use his wealth, power and social status to maximize his reproductive fitness.

The dynamic process of cultural selection will stop only if it reaches a medium-run Nash equilibrium point. At that point, some set of memes is established in the host population, and these memes are in effect dictating to their hosts an evolutionary stable strategy (possibly a mixed strategy) that facilitates the replication of the memes. To sustain an efficient equilibrium at which nobody can be made better off in terms of cultural fitness without making somebody else worse off, people must believe that they will be suitably punished if they deviate from the prevailing cultural norms and conventions. Rational agents still have no reason to cooperate with genetic 'strangers' in the long run, unless threatened with sufficiently harsh punishment such as severe physical injury or confinement. In the medium run, however, the threat of milder and more informal forms of punishment may help to make these agents comply with cultural rules that, while biased in favour of the selfish concerns of successful members of society, supplement, hide and even modify to some degree the biological imperatives. Social stigma, disapproving looks, and other warning signals might even provide enough help to sustain a Nash bargaining equilibrium in a large anonymous society in the medium run.

But controversy remains about whether the costs of detecting and punishing deviants, including the second-order costs of punishing those who fail to report non-cooperation by others, would be prohibitive in large societies in which most people are strangers. These costs might be manageable if disapproving looks and so forth could suffice to prevent cheating among strangers who have internalized the relevant memes at a medium-run equilibrium. Nevertheless, the issue remains alive since these agents apparently remain selfish at their core. If selfish genes remain predominant in determining rational behaviour even in the medium run, the milder forms of social punishment might not help much to sustain mutual cooperation. Moreover, the costs of establishing official agencies to detect these self-interested deviants and inflict the kinds of harsh legal punishment that may be needed to sustain cooperation among strangers, would certainly create substantial inefficiencies, as Binmore (2005, 82) admits. Widespread cooperation might even prove impossible in a large anonymous group.[7]

[7] For relevant discussion, see Gintis 2006, Ross 2006, and Seabright 2006, as well as the reply by Binmore 2006.

4. The Short Run

In the short run, also known as 'economic time', both 'biological time' and 'social time' are treated as frozen, and both genes and memes are regarded as parameters. The parametric memes include standards of interpersonal comparisons, which rational agents must rely on to make common assessments of what's fair. Viewed from the perspective of 'economic time', rational humans are consciously making their everyday market decisions in the context of given social customs and norms, keeping in mind that the decisions and customs are ultimately constrained by selfish genes seeking to replicate themselves. Rational humans are, it seems, predominantly selfish agents consistently acting to get more scarce resource for themselves and their close associates, subject to budget constraints.

Competitive markets are never out of short-run equilibrium for long. They are assumed to "adjust to an unanticipated piece of news ... in minutes or hours" (157). Efficient equilibria may be sustainable by threats of various forms of physical and social punishment for deviating from the given conventional rules and norms of market exchange. At the same time, memes and genes are evolving in the population on their respective time scales. The evolutionary processes are thus tending to integrate market behaviour within Nash bargaining equilibria in the medium run and long run, respectively.

Binmore also recognizes that humans are more or less prone to making errors about the games they are playing in the short run. They may still be in the process of learning which cultural norms and rules apply to some unanticipated situation, so that they make strategy choices on the mistaken assumption that they are playing one game whereas in fact they are playing another. People might choose temporarily to cooperate in a prisoners' dilemma game, for instance, until they learn that their traditional cultural norms of cooperation do not apply to such a game. To the extent that agents are temporarily confused in some such way, Binmore ignores their short-run behaviour as mere 'noise'.

5. The Nature of Human Nature

If the game of life has the complex structure of biological, cultural and economic time horizons as indicated, then a potential justification emerges for why rational agents in any social context should be assumed to be predominantly selfish. Given that his genes in their drive to replicate themselves are ultimately determining his behaviour, a rational human simply has no alternative but to consistently look out for himself and his kin: he not only will do so but he also *should* do so in the final analysis. Some support for this speculation can be gathered from observation of human behaviour. Even if people are observed to behave consistently as *if* they are selfish genes, however, such empirical evidence alone cannot justify the claim that people are biologically determined to behave this way. To justify the claim, biological science must identify and test a causal mechanism in terms of which a rational human's genes are seen to determine his behaviour.

Given such a mechanism, the 'selfish gene' metaphor is more than just a metaphor: the rational person's genes can be validly inferred to instruct him to make the relevant strategy choices. Moreover, since all rational humans are ultimately motivated by such biological considerations, cultural norms will inevitably tend to reflect the self-interest of those perceived to be successful in any given social context. No doubt genes do not provide instructions for strategy choices in every situation. But they remain sufficiently influential, it seems, to constrain cultural norms to embody selfish practices of successful individuals. Rational agents will seek wealth, power, fame, and other advantages for themselves and their families (as well as a few friends regarded as family) in accord with the prevailing social conventions. The conventions themselves will encourage and permit such self-interested behaviour. Everyone cooperates on that understanding, provided they expect that breaking the rules will provoke suitable punishment by others.

The biological argument that rational agents will invariably choose to replicate their genes deflects the objection that Binmore provides no justification for the assumption that rational agents are at bottom selfish. The objection would have force if the axiom of selfishness were simply derived from observing people's behaviour. In that case, the objectors may say, selfish behaviour is not inevitable but merely the product of some social institutions rather than others. Since it is merely a cultural artifact, selfishness might eventually be removed by suitable social reforms. In Binmore's framework, though, there is arguably a reason why every rational agent *ought* to pursue his selfish concerns, why he *ought* to cooperate to mutual advantage with other selfish agents, and why he *ought* to coordinate with them on a fair outcome that is necessarily a proportional bargaining solution. The reason is that it is impossible for rational agents to do any better than this, given the deep structure of human nature.

If this is right, then, for Binmore, the nature of human nature is such that rational humans are selfish at their core, even though their preferences may display a variety of contents depending on context. He can be said to hold a constrained version of ethical pluralism, according to which rational self-interested people pursue various irreducible goods in different settings. The pluralism of goods is ultimately constrained by the fact that people are at bottom selfish genes.[8]

[8] Binmore (43) argues that the 'naturalistic fallacy' is avoided by substituting hypothetical imperatives for categorical ones. Although it avoids the problem of deducing 'ought' from 'is', however, this substitution does not by itself imply that there are things which every rational person ought to find desirable because they are objectively good for all human beings. The substitution is compatible with cultural relativism, where by 'relativism' is meant a doctrine which holds that whatever a society conventionally reports as 'good' or 'fair' is good or fair for that society. Whatever a Nazi society traditionally endorses would then be good or fair for that particular society, despite the protests of Jewish outsiders who happen to observe the social practices. Although he endorses what he calls 'relativism', Binmore indicates that there are universal goods and bads. For instance, reproductive fitness, and to that extent self-interest, seems to be a universal good. 'Fairness' as he understands it also has a component which is built into the human genome, so that a proportional bargaining equilibrium is universally good whatever the cultural context. Nevertheless, it remains unclear that what I am calling his doctrine of constrained pluralism can avoid the objection that it admits a Nazi society as good or fair on its own terms. The problem is that universal goods and bads are confined to the deep biological structure of the game of life, and they may not constrain cultural content beyond

This assumption that people are ultimately selfish is a crucial feature of Binmore's framework, as I understand it, since it leads him to dismiss as 'utopian' certain moral and political arrangements, including utilitarian arrangements, which agents obsessed with their own selfish goals have no reason to sustain as equilibria.

6. The Game of Morals

Binmore argues that rational agents, consistently with their selfish nature, can play an indefinitely repeated 'game of morals' in the short run, where their everyday decisions are made. As unanticipated shifts of the feasible set of outcomes occur, these agents are able to make use of an innate moral sense to quickly distribute the social surplus or deficit by mutual consent, without needing to engage in costly and time-consuming face-to-face Nash bargaining. More specifically, any human has the moral capacity to imagine himself in an 'original position' under a 'veil of ignorance', where he forgets his own particular identity and puts himself with equal probability in the positions of every member of society to see things from each of their perspectives. When he is in another person's position, the agent empathizes with that other person in the sense that he imagines himself revealing the same personal preferences which that other person is known to have when in that position: the agent's hypothetical choices are supposed to mimic the other person's actual choices over possible outcomes. This moral ability to forget our own identities and put ourselves in another's place to experience his utility payoffs as if they were our own is, Binmore speculates, "built into our genes" and thereby forms "the deep structure" of our fairness norms (129).

To form an empathetic utility function, any rational agent empathizes with each member of society by adopting that member's personal utilities as his own while in that member's shoes, and then weighs how good he thinks it is to be one person with his personal utilities in comparison to another person with hers. He might decide that it is better to be Albert drinking a bottle of Montrachet, for instance, than it is to be Betty slaving away in the cotton fields, and that it is better to be Betty lifting a bale of cotton than himself facing conviction for a serious crime which he didn't commit. But to form such empathetic preferences, he must make interpersonal comparisons of utility. More importantly, for everyone to form the same empathetic preferences, there must be a consensus on a standard of interpersonal comparisons. But there is no presumption that different people could form the same empathetic judgments before entering the original position. Rather, different people will typically have conflicting empathetic preferences when they know their actual positions in society. They enter the original position to forget their identities and, when imagining themselves in another's shoes, they adopt his empathetic preferences. Given the conflicting empathetic preferences, each party must predict that the same implicit bargain

requiring it to reflect to some extent the selfish concerns of individuals who are perceived to be successful in terms of reproductive fitness.

or symmetric 'empathy equilibrium' will be achieved for everyone to agree to make the same empathetic judgments, keeping in mind that each party assumes that he has an equal probability of being any member of society when he leaves the original position. But for this to happen, they must agree on the standard of interpersonal comparisons which should be used in any given situation.

In Binmore's view, common standards of interpersonal comparisons are determined by social evolution. A meme that instructs its human hosts to make the relevant interpersonal comparisons becomes established in the population at a Nash bargaining equilibrium in the medium run. True, these cultural standards of interpersonal comparisons will more or less reflect the unequal bargaining power of the members of society. Yet everyone in the society accepts these cultural assessments of each person's 'worthiness' in comparison to another's at that equilibrium. As a result, everyone will regard a deal struck in the original position as fair if it makes use of these cultural standards. Indeed, a major reason for entering the original position seems to be to focus impartially on these cultural norms. Even selfish people will enforce the deal as fair once they emerge from the original position and return to their actual social positions. They have reason to enforce the deal because it involves empathetic assessments that embody the standards of interpersonal comparison which everyone already accepts as part of their culture. Since everyone is making the same empathetic judgments, everyone's empathetic utilities are equal.

As Binmore shows, a self-enforcing fair bargain of this sort must be a proportional bargaining equilibrium of the game of morals. To determine any person's worthiness relative to another's in a given situation, rational agents behave as if they are instructed by the relevant meme to set the ratio of the two persons' cultural indices of worthiness such that a proportional bargaining solution to the game of morals corresponds to a Nash bargaining equilibrium of the game of life in the medium run. Given an unanticipated change in the feasible set of outcomes, such culturally-determined ratios, which are fixed in the short run, can be used in the original position to quickly yield an equilibrium outcome which everyone regards as fair, without the need for actual face-to-face bargaining.[9] More precisely, any social surplus or deficit is quickly distributed such that any person i's utility gain or loss $(u_i - \hat{u}_i)$, weighted by his cultural index of worthiness w_i, is *equal* to that of any other person j:

$$\forall i, j : \frac{u_i - \hat{u}_i}{w_i} = \frac{u_j - \hat{u}_j}{w_j} \tag{1}$$

where u_i is person i's utility level after the distribution, and $\hat{u}_i$ is his utility level beforehand set at the given Nash bargaining equilibrium in the medium run. Elementary rearrangement of terms gives:

$$\forall i, j : \frac{u_i - \hat{u}_i}{u_j - \hat{u}_j} = \frac{w_i}{w_j} \tag{2}$$

[9] It may seem a bit unusual to suppose that cultural standards of interpersonal comparison have evolved and become established at a medium-run equilibrium *prior* to ever being used in the original position to generate a fair equilibrium in the short run. But rational agents may anticipate the usefulness of such standards before ever using them.

Evidently, the larger any person i's cultural index of worthiness is compared to another person j's in any situation, the larger is person i's share of any social surplus or deficit in comparison to person j's share. Any person's relative share can be cashed out in terms of "such parameters as need, effort, ability or status", depending on the interpretation of what counts as worthiness in the given context (179–83).

Fairness thus understood is a short-run phenomenon. It is used to quickly distribute any unanticipated social surplus or deficit before there has been time for cultural norms, including standards of interpersonal comparisons, to adjust to the new situation: "cultural evolution will often be too slow to keep up with the rate at which new coordination problems present themselves" (160). If unanticipated changes in the feasible set cease to occur for a sufficiently long time, however, cultural norms will have a chance to evolve to a new Nash bargaining equilibrium in the medium run. At this new equilibrium, fairness is leached out of people's behaviour because there has been no need to distribute any unanticipated social surplus or deficit for a period of some 'months or years'. But, while no longer visible, the notion of fairness apparently remains latent in the public culture. If unanticipated changes in the set of feasible outcomes start up again, rational agents can again make use of the original position device to coordinate quickly on a proportional bargaining equilibrium, although now with new standards of interpersonal comparison associated with the new Nash bargaining solution of the game of life in the medium run.[10]

Binmore (139–143) suggests that the capacity to imagine ourselves in strangers' positions, empathize with them, and apply cultural standards of interpersonal comparison to reveal common empathetic judgments, is a refinement of the ability to put ourselves in the shoes of our future selves as well as kin, sympathize with them, and add their personal utilities to our own utilities when making strategy choices. But he regards the distinction between empathy and sympathy as crucial (114–116). A person sympathizes with another by counting the other's personal utilities as his own, as when a person acts in accord with Hamilton's rule: the one's personal choice is influenced by his biological affinity or love for the other person and to that extent facilitates their mutual cooperation. In contrast, a person empathizes with another by imagining himself in the other's position to make the same choices over outcomes which the other would make *in that position*. Empathy does not imply sympathy for the other. The agent need not make those same choices when back in his own shoes. He has no reason to reconsider or revise his personal choices but instead employs cultural standards of interpersonal comparison to express a common empathetic preference. When expressing similar empathetic judgments, agents are simply endorsing the cultural norms that determine the worthiness of being in one per-

[10] By implication, the Nash bargaining equilibrium in the medium run always includes cultural standards of interpersonal comparison which can sustain an outcome deemed to be fair in the short run, even though the standards are not explicit in the medium-run behaviour of rational agents. As Binmore points out, parties in the original position "mustn't care who they turn out to be if the current social contract is continued by default because they are unable to agree [on a fair outcome] ... The current equilibrium is therefore always deemed to be fair." (173)

son's shoes making that person's choices as compared to the worthiness of being in another's position making the other's choices.

If strangers sympathized with one another like kin, then everyone might identify with each other so closely that all would make identical utilitarian personal choices over outcomes. In that case, utilitarian outcomes could be achieved as fair outcomes. But Binmore emphasizes that sympathy is confined to kin and a few intimate associates who are treated as kin. Strangers cannot be expected to feel so much affinity for one another that they will cease to reveal selfish personal preferences and instead reveal identical social preferences.[11] Rather, they will inevitably remain separate agents with their conflicting selfish concerns. They will at best empathize with each other in the original position, employing cultural standards of interpersonal comparison to coordinate on a proportional bargaining equilibrium which reflects their unequal bargaining power in society. Any 'utopian' attempt to implement a utilitarian solution rather than a proportional bargaining equilibrium will result in social disaster. According to the utilitarian solution, an unanticipated social surplus or deficit would be distributed so as to maximize the sum of weighted personal utilities after the distribution:[12]

$$\forall i, j : \frac{u_i}{w_i} + \frac{u_j}{w_j} \tag{3}$$

But this utilitarian bargain in the game of morals generally doesn't correspond to a Nash bargaining equilibrium of the game of life. Rational selfish agents will thus generally fail to implement the utilitarian outcome as fair in the absence of an external enforcement agency.[13]

Binmore's theory of fairness is sophisticated and worthy of careful study. It clearly combines many important insights from Hume, Rawls and Harsanyi into a distinctive neo-whig brew. Nevertheless, powerful objections can be raised against it, in my view, objections which Hume himself may well have endorsed.

---

[11] Binmore would therefore dismiss as unappealing Sen's (1970, 156) so-called "identity axioms" relating to sympathetic (or 'extended') preferences. This seems to be why he classifies Sen along with Mill as "the kind of do-gooder who thinks [he] knows better what is good for people than people do for themselves" (122). Yet his depiction of these great defenders of individual liberty and diversity is hardly judicious. He apparently assumes that Sen or Mill would force people to accept as fair, say, a utilitarian outcome. But there is no reason to suppose that either thinker would ignore the fact (when it is a fact) that the people don't feel the sympathy for each other upon which their voluntary implementation of such an outcome depends.

[12] Strictly speaking, there might also be a utilitarian outcome prevailing before the unanticipated change in the feasible set. But any previous utilitarian solution is irrelevant to the calculation of the new utilitarian outcome. Thus, we can speak of maximizing the sum of personal utility levels after the distribution, rather than of maximizing the sum of personal utility gains or minimizing the sum of utility losses from the previous utilitarian status quo.

[13] A utilitarian bargain might be sustained within a proper subset of society, Binmore admits, because members of society outside the subset enforce the bargain by threatening suitable punishment of cheaters. But a utilitarian outcome will not be sustained as a fair equilibrium by society as a whole.

7. Another View of Human Nature

There is continuing controversy over the use of the selfish gene paradigm to explain human behaviour.[14] But the weight of the argument appears to rest with the critics. Leading biologists such as Stephen Jay Gould and Richard Lewontin have rejected the paradigm as unscientific. Lewontin, for instance, dismisses sociobiology as 'ideology' for failing to specify, let alone test, any causal mechanism in terms of which genes could be shown to significantly constrain a person's behaviour:

> "[W]e need some appropriate mechanism of biological mediation between genes and behavior. Can there be genes 'for' being nicer to your brothers and sisters than to your second cousin once removed ...? Can genes really modulate the structure of the central nervous system to produce just the right contingent behavior? Nothing is known by way of an answer to this question, and, more important, there is no program of empirical work on the central nervous system intended to make these formal speculations into concrete anatomy and physiology." (Lewontin 2000b, 328–329)

Gould also argues that, although "the range of our potential behavior is circumscribed by our biology," there is "no evidence whatever" that specific genes determine specific human behaviours, including "reciprocal altruism" (Gould 1976, 16, 20).

Following John Alcock (2001), Binmore retorts that "critics like Lewontin or Gould ... pretend not to understand that sociobiologists seek explanations of biological phenomena in terms of *ultimate* causes rather than *proximate* causes" (6, emphasis original). An explanation of human behaviour in terms of proximate biological causes gives an account of the internal machinery whereby genes (in Lewontin's words) "modulate the structure of the central nervous system to produce just the right contingent behavior", whereas an explanation in terms of ultimate causes gives an account of the reproductive value of the behaviour. We may be ignorant of the proximate explanation, Binmore insists, yet we can still validly infer an ultimate explanation according to which rational humans behave as if they are instructed by their genes to maximize their reproductive fitness.

Nevertheless, the absence of a proximate biological explanation casts serious doubt on any ultimate explanation of human behaviour in terms of reproductive fitness because, without an account of the relevant internal machinery whereby genes produce the behaviour, we cannot tell if biology or culture is the more powerful determinant of the behaviour, or even if genes exert any significant influence. Even if human conduct is ultimately influenced by genes in some way, there is no evidence that the influence is ever anything but weak.[15] Any living organism is arguably "the nexus of a very large number of weakly determining

[14] See, for example, Maynard Smith 1995 and Gould 1997a; 1997b, among many other contributors to the debate surrounding Dennet's 1995 defense of the selfish gene paradigm.

[15] Moreover, if selfish genes exert a significant influence, it's puzzling that behaviours such

forces, no one of which is dominant" (Lewontin 2000a, 76). In the case of human beings, where cultural forces must be added to the mix and may possibly overwhelm biological forces, the claim that genes exert enough influence to determine specific behaviours seems especially shaky.[16]

As opposed to the selfish gene metaphor, a more reasonable scientific hypothesis is that human nature is highly plastic. If genes don't significantly constrain human behaviour, humans are capable of learning to behave in a wide variety of ways, only some of which are admirable and just. As Gould says, "the idea of biological determinism, with specific genes for specific behavioral traits", should be replaced by "the concept of biological potentiality, with a brain capable of a full range of human behaviors and predisposed toward none" (Gould 1976, 20, also quoted by Alcock 2001, 134). Indeed, he suggests that "most of our mental properties and potentials" may have arisen as nonadaptive "spandrels" rather than adaptations for reproduction:

> "The human brain is the most complicated device for reasoning and calculating, and for expressing emotion, ever evolved on earth. Natural selection made the human brain big, but most of our mental properties and potentials may be spandrels—that is, nonadaptive side consequences of building a device with such structural complexity." (Gould 1997b, 52)

Lewontin takes a similar line:

> "Our DNA ... makes possible the complex brain that characterizes human beings. But having made that brain possible, the genes have made possible human nature, a social nature whose limitations and possible shapes we do not know except insofar as we know what human consciousness has already made possible." (Lewontin 1991, 123)

Both Gould (1976, 22) and Lewontin (1991, 123) refer with approval to Simone de Beauvoir's maxim that "*a human being is 'l' être dont l' être est de n' être pas,*' the being whose essence is in not having an essence."

This Gould-Lewontin view of human nature implies that sociobiology, evolutionary psychology conceived in exclusively adaptationist terms, or "any other

---

as gay and lesbian lifestyles and the taboo against incest are apparently so widespread and persistent, even if they might disappear in the long run. For a critique of sociobiological explanations of gay behaviour, see, for instance, Lewontin 1991, 101–103.

[16] Alcock 2001, 12–16, 32–56, 149–187, emphasizes that sociobiology is perfectly compatible with explanations of human behaviour in terms of proximate causes, biological or cultural, even though its own focus is on explanations in terms of ultimate causes. Yet he doesn't clarify the relative weights of cultural and biological factors in any context, and never considers the possibility of genuine conflicts between them. Rather, he takes for granted the validity of ultimate biological explanations, and then suggests that biological imperatives properly constrain proximate cultural explanations (Alcock 2001, 130–131). In short, biology constrains culture, and any adequate explanation of human behaviour will involve some complementary mix of biological and cultural ingredients. Binmore seems to take a similar view, although he tends to associate it with virtually unlimited cultural 'relativism' whereas Alcock 2001, 134–147, tends to stress universal biological constraints on cultural variation.

ultra-Darwinian theory" is "a particularly dubious approach to human behaviour" (Gould 1997b, 52). The view might be called a 'blank slate theory', provided the theory is understood merely as holding that the human brain is empty of any concrete *a priori* concepts and categories, including any which might have been inherited through natural selection. Thus understood, a blank mental slate at birth is compatible with the claim that, so far as we can tell, the brain has certain capacities but not others. The brain registers only a limited range of sensations and emotions, for instance, and it experiences only some co-existences and sequences of natural phenomena even if it can imagine others never actually experienced. Although it may not be predisposed toward any specific behaviours, therefore, a rational brain acquires specific behavioural dispositions on the basis of its experience of the natural world. The brain is able to develop and transform itself in this fashion, from a blank slate into a rational apparatus with certain specific dispositions, by learning that nature is organized in accord with certain causal laws, and that reasonable concepts and categories of human good must take account of these natural regularities. Social institutions play a central role in this learning process, although learning doesn't always depend on imitation of others.[17]

My reliance on critics like Gould and Lewontin is predicated on the assumption that Binmore intends to make use of the selfish gene paradigm to justify the claim that rational agents will inevitably tend to exhibit selfish behaviour as they have time to adjust to their social circumstances. But no objection to the paradigm is required if he doesn't intend to use it for such explanatory purposes. He might use it merely as a bookkeeping tool, to record the fact that any given human population will contain some genes rather than others. The latter use is tautologous, since no attempt is made to explain behaviour in terms of genetic predispositions. On the off-chance that this is all he means to say, however, he has no biological reason to conclude that human nature is essentially selfish so he would presumably agree that observed selfish behaviour may simply be the product of some social institutions rather than others. Unfortunately, whatever he intends, Binmore's use of the selfish gene paradigm sends a social and political message that rational agents are inevitably selfish, and that any program of social and political reform which fails to recognize this ultimate fact of human nature is utopian.

The selfish meme paradigm for explaining culturally-determined behaviour is also problematic. It too can be employed in a tautologous way, as a mere bookkeeping tool rather than an explanatory framework.[18] Nevertheless, I wish to bracket any criticisms to argue that, if human nature is not innately selfish, self-replicating memes at an efficient Nash equilibrium point may instruct rational agents to choose strategies which everyone regards as fair even though fairness

[17] Alcock 2001, 134–147, is highly critical of what he takes to be 'blank slate theory', whereas Binmore 2005, 46, says that "genes preempt much of the space on our mental slates" yet "we are nevertheless born with large blank spaces for experience to fill". But even Binmore apparently wishes to maintain that genes significantly constrain how cultural evolution fills up the "large blank spaces" which remain on our mental slates at birth.

[18] Distin 2004 defends memetics whereas Richerson and Boyd 2004 offer criticisms and a distinctive approach to cultural explanation.

is not conceived in Binmore's terms, as a proportional bargaining solution of the game of morals.

8. Utilitarian Justice Reconsidered

Suppose that human nature is highly plastic, in which case rational agents might develop into predominantly social agents, with dispositions to mutually cooperate which are far stronger than their dispositions to cheat even when others don't threaten to punish them for cheating. Cultural evolution might produce such a result. Imagine, for instance, that a complex meme invades a population and instructs its human hosts to recognize and arrange into a coherent system certain basic individual rights and correlative obligations. These equal rights or claims, according to the meme's instructions, protect vital personal interests which every person shares by virtue of his human nature. Any rational agent infected by the meme recognizes the system of human rights as self-evident, and assigns far more importance to these self-evident rights and duties than to any of his competing selfish concerns. In short, individuals learn to reciprocate in terms of a code that distributes the basic rights: each agent learns to respect others' rights in return for others' respect of his, and each learns to expect harsh punishment in return for his failure to satisfy his correlative duties to others. An efficient equilibrium of the game of life exists in which each person's strategy choice is suitably correlated with others' strategy choices: nobody has any incentive to alter his respect for others' rights given that the others are choosing to respect his.

But beyond learning to play culturally-determined strategies of mutual cooperation in terms of equal rights, rational agents can also come to identify themselves with such a strategy, to see it as an extremely valuable part of their own plastic nature. In other words, they view the social feeling of being in unity with their fellows with respect to basic rights as a natural feeling which, though acquired and developed on the basis of experience, is more important to their self-conception than any conflicting feelings, including selfish desires to violate others' rights. As a result, they conceive of themselves as social and moral agents with a powerful natural desire to cooperate in terms of basic rights, not as selfish agents with a dispensable veneer of morality that merely facilitates in the short run what face-to-face Nash bargaining would bring about anyway in time. Their natural desire to cooperate gives rise to powerful moral feelings that any person deserves severe punishment if he fails to comply with the social rules of justice which distribute equal rights and duties, since non-compliance endangers mutual cooperation and the common good. Every person feels that he ought to be punished by himself even if not by others for his failure to satisfy his obligations correlative to others' basic rights. Thus, in the course of developing his desire to cooperate in terms of equal rights, he develops a conscience, that is, an internal monitor that observes his own behaviour and inflicts harsh retaliation in the form of suitably intense feelings of guilt for any violation of another's rights. This "judicious spectator", as Hume (1888a, 581) calls it, is not tied logically

to an omniscient and omnipotent god. Nor is it necessarily tied to an innate 'moral sense'. Rather, it amounts to an habitual desire to do right and avoid doing wrong, that is cultivated in every rational person's 'heart' (that is, at the core of his nature) through cultural evolution.

Any rational agent who acquires such a conscience still places himself in an original position under a veil of ignorance to determine what's fair. But now the situation is not properly modeled as a 'game of morals' played by distinct agents who rely on cultural standards of interpersonal comparisons to arrive at a proportional bargaining equilibrium. Rather, the situation is not really a game at all but a one-person decision problem. The whole point of the exercise is to ignore one's own particular circumstances so as to recall and focus attention exclusively on the basic rights which, according to the culture, every person possesses. An agent immersed in that culture believes that he acts fairly as long as he respects others' rights, and he can remind himself what are considered basic rights through solitary reflection in the original position. True, he still imagines himself as any member of his society with equal probability, revealing each person's preferences while in that person's shoes. Yet he now always sees himself as the same human being when in any person's shoes, making the same basic claims and satisfying the same correlative obligations. In effect, he sympathizes with that one human being while in every person's social position, and arrives at fair decisions by maximizing the sum of the unweighted personal utilities associated with behaving in accord with social rules of justice that distribute the basic rights and duties:

$$\forall i, j : u_i + u_j \tag{4}$$

This utilitarian solution implicitly relies on cultural standards of interpersonal comparisons: all persons are deemed to be equal—the same human being—in terms of the vital personal interests that, according to the culture, ought to be treated as basic rights. The solution still generally doesn't correspond to a Nash bargaining equilibrium of the game of life. But now there isn't any need for it do so. The utilitarian outcome corresponds to an efficient Nash equilibrium point, or cultural evolutionary stable strategy, which need not be a Nash bargaining solution.[19]

[19] Binmore's selection of the Nash bargaining solution from among the other possible candidates for an efficient equilibrium of the game of life is tied to how players in the game of morals can reasonably be expected to bargain in the original position, where the assumption of equal bargaining power makes sense "because the available bargaining strategies behind the veil of ignorance are the same for [all] players" (26, note 4). Given that the players have equal bargaining power, and that "the characteristics of the players and the nature of the bargaining problem are common knowledge, ... a number of different lines of enquiry then converge on identifying the *Nash bargaining solution* as the rational agreement" (25–26, original emphasis). Equal bargaining power in the original position is compatible with unequal bargaining power in the game of life. The unequal power is brought into the game of morals through the cultural standards of interpersonal comparisons which are employed in the original position to arrive at a proportional bargaining equilibrium that corresponds to a Nash bargaining equilibrium of the game of life. Skyrms 1996, 107, agrees with Binmore that the Nash bargaining solution deserves "more attention" in any philosophy of distributive justice. But Skyrms 1996, 108,

The basic rights identified as obviously belonging to any person by virtue of his humanity might not be recognized before an advanced stage of cultural evolution. They might also vary in content to some extent across different advanced cultures, although at least some overlap is to be expected in light of common biological vulnerabilities. In any event, these self-evident human rights can only be expressed in rather abstract and vague terms, as when Hume speaks of the basic right of any man to control the fruits of his own 'art or industry', where control includes keeping the fruits or bestowing them on others. Some such basic right, he insists, is obviously in the general interest:

> "Who sees not, for instance, that whatever is produced or improved by a man's art or industry ought, for ever, to be secured to him, in order to give encouragement to such *useful* habits and accomplishments ... Examine the writers on the laws of nature; and you will always find, that, whatever principles they set out with [including a natural right to property], they are sure to ... assign as the ultimate reason for every rule which they establish, the convenience and necessities of mankind ... [P]ublic utility is the *sole* origin of justice, and ... reflections on the beneficial consequences of this virtue are the *sole* foundation of its merit." (Hume 1888b, 183, 195, original emphasis)

Yet this basic right to control the fruits of one's productive efforts is ambiguous as it stands. It doesn't even distinguish between capitalism and socialism, for example, since it's compatible with either individual ownership of the means of production or a community of producers mutually consenting to pool the fruits of their efforts and distribute the total fruits according to some fixed equitable principle such as perfect equality or "from each according to his ability, to each according to his needs". Indeed, it's compatible with a mixed economy where capitalistic partnerships or corporations, involving individual owners each of whom is entitled to control a share of profits in proportion to his capital investment, compete in the same markets with socialistic cooperatives, in which there are no individual owners.[20]

To remove the ambiguity surrounding any such basic right, any given society requires a legislative process to effectively transform the abstract basic rights into concrete positive rights: positive laws are needed to specify in detail that particular community's interpretation of the basic rights, supplement the basic rights with auxiliary rights and other legal instruments such as powers and immunities which may be needed to carry the basic rights into effect, and so forth. Without arguing the point here, any such legislative process should be democratic in form, with the important caveat that a citizen who participates is properly conceived as a moral agent rather than a selfish strategic actor. A citizen is properly seen as a person who is constrained by his culture to consider

---

also points out that cultural evolution can generate perfectly correlated beliefs and behaviours, resulting in a utilitarian equilibrium outcome.

[20] I don't mean to imply that Hume imagined these possibilities, although he was acquainted with Rousseau, who did conceive the possibility of socialism.

as genuine positive laws only rules that distribute equal concrete rights, where the concrete rights are, in the legislature's estimation, the best interpretation for that community of underlying basic rights that, according to the culture, self-evidently belong to every human being. The legislative process on this view is not a game played by rational agents with distinctive individual goals but rather an epistemic device for generating a maximum likelihood estimate of the best concrete rules of justice for that community.[21] Every rational participant in the legislative procedure shares one goal, namely, the discovery of equal concrete rights reasonably held to promote the common good, which has an objective component in that culture, that is, the basic abstract rights due to everyone. Every participant properly sees himself as the same human being guided by the same cultural norms, seeking to find perfectly correlated strategies which every rational agent in his society ought to endorse to achieve social justice.

This picture of a rational agent as somebody who is motivated ultimately by social and moral feelings, who sympathizes and chooses to be at one with his fellows by coordinating with them on the same system of rights and correlative duties, may be an ideal picture. But it is by no means utopian in the sense of being impossible in light of human nature. Indeed, such an agent, though he would feel intense guilt if he failed to satisfy his conscientious desire to respect others' rights, remains free to pursue his selfish concerns where permitted by his own equal rights. Though he conceives of himself as ultimately in unity with his fellows, therefore, there is still room for him to engage in selfish behaviour to a limited extent. He can engage in capitalistic market behaviour, for instance, given that his society recognizes rights of private property and contract. Moreover, broad inequalities of income, wealth, social status and other primary goods will be viewed as compatible with distributive justice in his culture, if his claims to help and relief entitle an individual to little more than subsistence even when disabilities prevent him from participating in the market. In contrast, other societies may establish decentralized socialistic economies together with highly egalitarian distributions of resources. There may well be less scope for selfish behaviour in such cultures. But nothing in the present approach prevents different societies from endorsing different concrete codes of justice and equal rights.

Much more needs to be said to clarify this utilitarian alternative to Binmore's neo-whig theory of justice.[22] But, for now, it suffices to say that Hume seems to have favoured some such utilitarian theory of justice. Admittedly, he argues that rules of justice originate in "self-love" and that "allowance for a certain degree of selfishness" must be made "because we know it to be inseparable from human nature, and inherent in our frame & constitution" (Hume 1888a, 583). But this doesn't imply that people are always predominantly selfish, or that they will inevitably try to cheat if the only thing stopping them is their own social and moral feelings: "a certain degree of selfishness" is not necessarily an overwhelming degree, and a limited degree of self-interested behaviour is permitted by a code of equal rights and duties. Hume insists that rational agents

[21] For discussion of Condorcet-type voting procedures as maximum likelihood estimation procedures, see, especially, Young 1988; 1995.

[22] For further relevant discussion, see Harsanyi 1977; 1992; and Riley 2005; 2006a; 2006b.

can cultivate the moral sentiment of justice to such a powerful pitch through the mechanism of sympathy for all other human beings that it 'overpowers' any contrary feelings of self-love:

> "We are naturally partial to ourselves, and to our friends, but are capable of learning the advantage resulting from a more equitable conduct ... And as the benevolent concern for others is diffused, in a greater or less degree, over all men, and is the same in all, it occurs more frequently in discourse, is cherished by society and conversation, and the blame and approbation, consequent upon it, are thereby roused from that lethargy into which they are probably lulled, in solitary and uncultivated nature. Other passions, though perhaps originally stronger, yet being selfish and private, are often overpowered by its force, and yield the dominion of our breast to those social and public principles." (Hume 1888b, 188, 275–276)

The passion for justice, though an 'artificial virtue' in the sense that (unlike self-love) it must be cultivated after reflecting upon its consequences for the common good, is not unnatural in the sense of being foreign to human nature: "Tho' justice be artificial, the sense of its morality is natural ... [A] sense of morality is a principle inherent in the soul, and one of the most powerful that enters into the composition." (Hume 1888a, 619) Justice has a foundation in an innate 'moral sense', an instinctive benevolence or sympathy for other humans, an original "sentiment of pleasure or pain, which arises from characters and actions, of that *peculiar* kind, which makes us praise [those characters and actions which conduce to the well-being of other humans] or condemn [those which are harmful to other humans]" (Hume 1888a, 472, original emphasis). This original "sentiment of humanity" can be "roused" to such a pitch through education and learning that rational agents eventually come to believe that it's natural for them to mutually cooperate in terms of a code of equal rights and duties, and unnatural—beneath human dignity—to cheat.[23]

9. Conclusion

As I understand it, Binmore's neo-whig theory of justice is embedded in a complex 'game of life' operating on three distinct time scales, to wit, biological time,

[23] Many commentators agree with Selby-Bigge that "the system of morals in the Enquiry is really and essentially different from that in the Treatise" (1888b, xxiii). Although I don't agree, it seems fair to say that Hume tends in the Enquiry to emphasize the extreme importance of the moral "sentiment of humanity", whereby one man sympathizes with others by virtue of their common human nature, as the foundation of justice. By contrast, he tends in the Treatise to highlight the constraints imposed by self-love on sympathy and benevolence. But this difference of emphasis can be reconciled once we see that sympathy may not extend much beyond a system of reciprocal rights, and that a degree of self-love is compatible with the individual's equal rights. In the end, Hume seems to think that the virtues of morality and justice are invariably in harmony with enlightened self-interest. See, also, Hume 1888b, 270–276, 281–284.

cultural time, and economic time. He argues that justice is manifest in a proportional bargaining equilibrium of a 'game of morals', which corresponds to a Nash bargaining equilibrium in cultural time of the game of life. His argument seems unassailable if rational agents are predominantly self-interested, an assumption that he is apparently willing to make on the grounds that human behaviour is ultimately constrained in accord with the selfish gene paradigm. But there is no compelling scientific evidence for that paradigm. Rather, human nature appears to be highly plastic. If so, rational agents might eventually be moulded by cultural forces into social and moral actors who effectively believe that they are the same person—no different from anyone else—when it comes to certain vital personal interests which ought to be treated as rights. Such people will mutually cooperate in terms of a code of equal rights and correlative duties, even if they are not threatened with punishment *by others* for failing to cooperate. Compliance with the rules is enforced by the actor's own conscience, a powerful internal 'judicious spectator' which threatens to inflict harsh punishment in the form of intense feelings of guilt for cheating.

Since there is no compelling reason to suppose that rational people are constrained by nature or culture to be predominantly selfish, there is no compelling reason to ignore the possibility that a utilitarian outcome, involving some distribution of equal rights and duties whose content is held to be extremely valuable for every person's wellbeing, might be generated by social evolution as an efficient equilibrium of the game of life. Indeed, there is textual evidence that Hume himself took seriously this possibility, even though he also believed that humans have an innate moral sense and that a degree of selfishness is inseparable from human nature.[24]

Bibliography

Alcock, J. (2001), *The Triumph of Sociobiology*, Oxford
Binmore, K. (1994), *Playing Fair*, Cambridge/MA
— (1998), *Just Playing*, Cambridge/MA
— (2005), *Natural Justice*, Oxford-New York
— (2006), Why Do People Cooperate?, in: *Politics, Philosophy & Economics 5*, 81–96
Dawkins, R. (1976), *The Selfish Gene*, Oxford
— (1982), *The Extended Phenotype: The Gene as the Unit of Selection*, San Francisco
Dennet, D. C. (1995), *Darwin's Dangerous Idea: Evolution and the Meanings of Life*,

[24] Some might even argue that a unique system of human rights is essential for maximizing the general welfare conceived in transcultural terms. The basic rights belonging to all persons, independently of cultural setting, might be grounded on a hedonistic psychology held to be applicable to all human beings, for instance, or on universal principles of rational choice according to which all humans should make the same interpersonal comparisons of utility at least when it comes to vital personal interests which all share. I haven't defended any such strong claim here, since I've left the basic rights to be culturally determined. But it is worth noting that Hume 1888a; 1888b and Mill 1969 apparently opt for some version of hedonism, whereas Harsanyi 1977; 1992 is committed to a strong version of rationalism which Binmore, 151, refers to as "the Harsanyi doctrine". Some cultural variation in codes of concrete rights is compatible with these approaches, given that different societies face different circumstances.

New York
Distin, K. (2004), *The Selfish Meme: A Critical Reassessment*, Cambridge
Gintis, H. (2006), Behavioral Ethics Meets Natural Justice, in: *Politics, Philosophy & Economics 5*, 5–32
Gould, S. J. (1976), Biological Potential vs. Biological Determinism, in: *Natural History 85 (5)*, 12–22
— (1997a), Darwinian Fundamentalism, in: *New York Review of Books 44 (10)*, 34–37
— (1997b), Evolution: The Pleasures of Pluralism, in: *New York Review of Books 44 (11)*, 47–52
Hamilton, W. D. (1964), The Genetical Evolution of Social Behavior, in: *Journal of Theoretical Biology 7*, 1–52
Harsanyi, J.C. (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge
— (1992), Game and Decision Theoretic Models in Ethics, in: R. J. Aumann/S. Hart (eds.), *Handbook of Game Theory.* Volume 1. Amsterdam, 669–707
Hume, D. (1888a), *A Treatise of Human Nature* (edited by L.A. Selby-Bigge), Oxford (second edition 1978, Treatise appeared first 1739–40)
— (1888b), *Enquiries* (edited by L.A. Selby-Bigge), Oxford (third edition 1975, Enquiries appeared first 1748–51)
Lewontin, R. C. (1991), *Biology as Ideology: The Doctrine of DNA*, Toronto (expanded edition 1993)
— (2000a), *The Triple Helix*, Cambridge/MA
— (2000b), *It Ain't Necessarily So: The Dream of the Human Genome and Other Illusions*, New York (second edition 2001)
Maynard Smith, J. (1995), Genes, Memes & Minds, in: *New York Review of Books 42 (19)*, 46–48
Mill, J. S. (1969), Utilitarianism, in: J. M. Robson (ed.), *Collected Works.* Volume X, Toronto/London, 203–310 (Utilitarianism appeared first 1861)
Rawls, J. (1971), *A Theory of Justice*, Cambridge/MA
— (1993), *Political Liberalism*, New York
Richerson, P. J./Boyd, R. (2004), *Not By Genes Alone: How Culture Transformed Human Evolution*, Chicago
Riley, J. (2005), Rousseau, The Social Contract, in: J. Shand (ed.), *Central Works of Philosophy*, Volume 2. London, 193–222.
— (2006a), Liberal Rights in a Pareto-optimal Code, in: *Utilitas 18*, 33–52
— (2006b), *Justice as Higher Pleasure.* Keynote Lecture, Mill Bicentennial Conference, University College London, April 5. (To be published in a volume edited by P. J. Kelly/G. Varouxakis, Cambridge)
Ross, D. (2006), Evolutionary Game Theory and the Normative Theory of Institutional Design: Binmore and Behavioral Economics, in: *Politics, Philosophy & Economics 5*, 51–79
Seabright, P. (2006), The Evolution of Fairness Norms: An Essay on Ken Binmore's Natural Justice, in: *Politics, Philosophy & Economics 5*, 33–50
Sen, A.K. (1970), *Collective Choice and Social Welfare*, San Francisco
Skyrms, B. (1996), *Evolution of the Social Contract*, Cambridge
— (2003), *The Stag Hunt and the Evolution of Social Structure*, Cambridge
Taylor, P./Jonker, L. (1978), Evolutionary Stable Strategies and Game Dynamics, in: *Mathematical Biosciences 40*, 145–156
Zeeman, E. C. (1980), Population Dynamics from Game Theory, in: Z. Nitecki/C. Robinson (eds.), *Global Theory of Dynamical Systems*, Berlin, 471–479

Young, H. P. (1988), Condorcet's Theory of Voting, in: *American Political Science Review 82*, 1231–1244

— (1995), Optimal Voting Rules, in: *Journal of Economic Perspectives 9*, 51–64