Nordic Journal of Political Economy

Volume 23 1996 Pages 151-173

Evolution of Fairness Norms

Ken Binmore

Ken Binmore *

Evolution of Fairness Norms

Harsanyi (1977) and Rawls (1972) are among those who have sought to capture the notion of fairness in terms of the bargain that would be freely negotiated behind a veil of ignorance that conceals the roles that the bargainers currently occupy in society. Rawls (1972) refers to the impartial standpoint generated by the veil of ignorance as the *original position.* This paper simplifies the study of the original position by restricting attention to the case of two bargainers, Adam and Eve.

Binmore (1994, 1996) criticizes the Kantian defenses of the original position proposed by Harsanyi (1977) and Rawls (1972) in favor of the more mundane view that the idea appeals to our moral intuition because we *already* use fairness norms for which the device of the original position serves as a stylized representative in settling picayune issues in our daily lives. Such an approach involves taking a naturalistic view of ethics within which moral rules are seen as devices for coordinating human behavior on an equilibrium in the game of life. In particular, fairness norms serve as equilibrium selection devices that allow a society to coordinate quickly on a Pareto-efficient equilibrium without damaging internal conflict.

Currently, we use fairness norms effectively only when settling small-scale questions like whose turn it is to wash the dishes, but the potential exists to employ the same principles on a grander scale. However, such a project has no hope of success unless we understand the circumstances under which successful fairness norms currently operate. In particular, we need to study how and why fairness norms evolved in the past in the hope of gaining insight into how society would adapt to attempts to use them systematically in solving large-scale coordination problems in the future.

This paper offers a speculative history for the evolution of fairness norms which respects the Linnean principle that 'Nature makes no jumps'. The proposed history begins with the primitive food-sharing arrangements that anthropologists sometimes

* Economics Department, University College London, UK
An expanded version of this paper will appear as Chapter 2 of Just Playing: Game Theory and Social Contract II, to be published by MIT Press. The support of the ESRC Centre for Economic Learning and Social Evolution is gratefully acknowledged.

propose as marking the beginnings of human sociality, and explains how we might have come to hold the empathetic preferences that are necessary as inputs when the device of the original position is employed. The approach developed in Binmore (1994, 1996) sees such empathetic preferences as being determined by social evolution. However, the paper leaves off at the point at which the problems raised by this thought begin.

Evolutionary ethics

Evolutionary ethics is widely misunderstood by moral philosophers. For example, G.E. Moore (1902) is still quoted as an authority for the claim that evolutionary ethics necessarily maintains that "we ought to move in the direction of evolution simply because it is the direction of evolution". It is then a short step to misrepresenting evolutionists as asserting that it is Right that the weak should go to the wall because only the fit should survive. Some Social Darwinists of Victorian times held such views, but modern social Darwinists like myself do not even believe that the teleological questions to which Moore thinks ethics should provide answers make any sense.

Evolutionary ethics denies that human societies exist to fulfil some great purpose. Evolutionists simply seek to understand why some types of human organization survive better than others. We may use this information in arguing that some reforms have a better chance of bringing about a lasting change in a society than others, but evolutionary ethics offers no authority whatsoever to those who wish to claim that some moral systems are somehow intrinsically superior to others. Evolutionary ethics is not a theory of the Good or the Right; it is a theory of the Seemly – and the seemly in one society may very well be unthinkable in another. Herodotus's story of the Greeks and Indians brought together at the court of Darius to compare their funeral customs says everything that needs to be said on this subject. The Greeks were horrified to learn that the Indians ate their dead fathers. The Indians were no less horrified to learn that the Greeks burned theirs!

Such frank moral relativism leads some critics to complain that evolutionary ethics is entirely empty. As de Laguna (1965) puts it, "It has been demonstrated again and again that the Darwinian theory will lie down with almost any variety of ethical faith." Of course it will! How could an evolutionary explanation of the origin of ethical systems fail to be consistent with any ethical system that has actually evolved? Like everybody else, social Darwinists often do not *like* the mores of societies with aspirations very different from their own, but evolutionary ethics teaches us not to label a society as Wrong or Bad according to some supposedly absolute standard.

The basic misunderstanding is that traditionalists think that to refuse to label a Society as Wrong or Bad is to say that all societies are equally Right or Good. But a relativist finds no more meaning in the claim that two societies are equally Good than he does in the claim that one is Better than the other. The incomprehension deepens when traditionalists hear relativists like myself expressing their aspirations for the kind of society in which they would like their children to live. Where do we get our opinions from? What is the source of our authority? Why do we bother at all if "nothing matters"?

The answers are simple. Like everyone else, social Darwinists are just mouthpieces for the memes[1] that have successfully replicated

1. This is Dawkins' (1976) term for the social equivalent of a gene – an idea, a rule-of-thumb, a norm, or anything else that can be replicated from one head to another by imitation or education.

themselves into our heads. We seek to replicate these memes into other heads because memes that don't induce such behavior in their hosts die out. Being relativists, we know perfectly well that other memes would be using us as instruments for their replication if we lived in different times or places. In ancient Athens, I would probably have chased after adolescent boys like Socrates. In antebellum Virginia, I would probably have been ready to keep slaves like Thomas Jefferson. But why should the fact that social Darwinists are willing to live with such obvious facts be thought to disqualify us from advocating reform? The opinions of those who claim the moral high ground are no less artefacts of their culture than ours. Nor are the underlying reasons that they seek to convert others any different.

When advocating reforms, traditionalists differ from social Darwinists only in claiming a spurious authority for their views. In seeking to persuade others, social Darwinists are constrained by the neodarwinian meme to making pragmatic appeals to the enlightened self-interest of our fellow citizens. But traditionalists tell themselves and others much more exciting stories in which our mundane arguments are trumped by appeals to Moral Intuition or Practical Reason or some other modern successor to Mumbo-Jumbo and the other gods who served as sources of authority for our ancestors. Sometimes, as when traditionalists invoke the Sanctity of Life as a reason for inflicting suffering on babies who have yet to be conceived, it is tempting to fight fire with fire by inventing new sources of authority, like Humanism or Gaia or Science – but the meme that is pushing this pen wants so much to be replicated that it won't even let me appeal to Intellectual Honesty in arguing against the creation of such new graven images.

Sociobiology and reductionism

Critics of social Darwinism who find the notion of a nonteleological ethics inconceivable, are only half the story. One must also beware of the enemies of reductionism, who wilfully misrepresent evolutionary theories in order to denounce them as simplistic. Such critics argue that evolutionary ethics is an attempt to reduce human morality to crude biological imperatives. So successful has this campaign been that one must now refer to much of what used to be called sociobiology as evolutionary psychology or behavioral ecology if one hopes to evade knee-jerk abuse. (See Tooby (1987) or Barkow *et al* (1992) for an account of the genuinely distinct features of evolutionary psychology. For behavioral ecology, see Barash (1982), Hughes (1988) or Krebs and Davies (1984).) But not even the most reductive sociobiologist ever held that evolution has written the rules of correct moral conduct into our genes. Sociobiologists believe in *co*evolution – according to which our biologically transmitted capacities and propensities evolved *in tandem* with the socially transmitted customs and conventions that govern primitive societies (Wilson, 1975, Lumsden and Wilson, 1981). In brief, the genes and memes relevant to prehistoric moral behavior evolved *together*, and so cannot sensibly be studied in isolation.

The evolution of language seems a particularly compelling example of this phenomenon. Everybody knows of Chomsky's (1959, 1965, 1980a) claim that all human languages share a common deep structure. Pinker's (1994) magnificent *Language Instinct* puts the case for the stronger claim that an innate capacity for language is actually carried in our genes. This is not to argue that Frenchmen carry genes for speaking French or that an American baby adopted by Japanese parents would find it any harder to learn Japanese than the natural children of his adoptive

parents. Just as our immune system is not just a stockpile of specific antibodies, but a piece of biological hardware that enables our bodies to create antibodies as and when required, so the language instinct is a hardwired learning device that makes it relatively easy for toddlers to learn languages structured according to certain innate principles. But this mental hardware is very far from tying down all the details of a language. On the contrary, French and Japanese are such different languages because our genes leave a great deal to be determined by *social* evolution.

Having made a powerful case for the existence of a human language instinct, Pinker (1994:420) joins Cosmides and Tooby (1994) and numerous others in speculating about the extent to which the *tabula rasa* theory of human nature is mistaken on other counts. His claim that the human sexual drive is innate is presumably uncontroversial. He is also doubtless right in asserting that we are genetically programmed to shun contaminated foods, or to flee from danger in certain situations, or to turn and fight in others. The robotic character of our behavior when in the grip of such strong emotions as love, disgust, fear or anger seems too transparent for serious doubt to arise on such questions.

It is more controversial to suggest that at least some of our capacity to empathize with our fellow men is instinctive. Plato defined man as a broad-nailed, featherless biped with a gift for politics – and modern studies of chimpanzee or baboon societies would seem to confirm that Plato was right to bracket our status as a political animal with the genetically determined characteristics of our anatomy. Critics of sociobiology take this claim to mean that we are genetically programmed to operate some *specific* political system. But nobody believes that our genes are programmed with any particular political political creed, any more than that we are programmed to call an apple a *pomme* or stack all the verbs at the end of a sentence. Like the language instinct, the political instinct operates at a deep structural level. As Hume (1939) explains, it is natural that our societies should be governed by "natural laws", but the specific "natural laws" that operate in a particular society are not determined by Nature. What moral philosophers call "natural laws" are either human inventions or else the product of *social* evolution.

I have been stressing the importance of the concept of coevolution to sociobiology, in order to prepare the way for a discussion of the most controversial of the items on Pinker's (1994:420) list of instinctive human traits. I think he is probably right to suggest that we not only come into the world equipped with an intuitive grasp of the essentials of human psychology, but that our sense of justice is also hardwired. This is not to say that our genes determine what will be regarded as fair in any particular society, only that they determine or constrain the algorithm that a society uses in deciding what is fair. But such an algorithm cannot operate without some input to chew on. I suspect that the necessary inputs are almost entirely socially determined. Since different societies have different social histories, it follows that they will make different fairness judgments. (Elster (1992) and Young (1994) see different fairness algorithms using similar social inputs being used in different contexts. My guess is rather that similar fairness algorithms are being used with different social inputs.) Even if Pinker is correct about the existence of a justice instinct, it therefore does not follow that ethics can be reduced to biology. On the contrary, with some important exceptions, I believe that almost everything in dispute among moral philosophers is

determined by *social* evolution.[2]

This paper seeks to clarify where the boundaries lie between biological evolution and social evolution in the model I use to study in social justice in a forthcoming book, *Game Theory and the Social Contract.*[3] I know perfectly well that the considerations that lead to the model are highly speculative and that the model itself is much too simple to come near capturing the richness of the human predicament, but the subject seems to me too important to wait until better founded and more sophisticated models become possible.

Evolution and justice

The early part of the road to an understanding of the evolutionary origins of human morality is by now so well-trod that there seems little point in attempting to do more than indicate some of the signposts along the way. Darwin (1971) and Huxley are more than a little lame on the evolution of ethics, but modern pathfinders like Alexander (1987), Boyd and Richerson (1985), Hamilton (1963, 1964), Maynard Smith (1982), Trivers (1971, 1985), Williams (1974) and Wilson (1975) have stories to tell that are theoretically coherent and strongly supported by the evidence from the field. However, those new to the subject might do better to begin with the more popular books of Dawkins (1976), Dennett (1995), Cronin (1991), Ridley (1993) or Wright (1994).

Paradoxically, it is only when such pilgrims on the evolutionary road get near their ultimate destination of modern man that their path begins to degenerate into a weed-strewn track of uncertain direction. It seems that we can observe other species and primitive human societies with a dispassionate objectivity that eludes our grasp when we turn our attention upon ourselves. However, just as ontogeny is said to recapitulate phylogeny, so I think that there is hope of sorting out our confused intuitions about human morality by tracing their origins back to their humble beginnings. Rather than hopelessly squaring the circle of human cooperation as in Binmore (1994, Ch. 3), we need instead to follow the title of Singer's (1980) pioneering work on evolutionary ethics by *Expanding the Circle.*

Singer traces the expansion of human moral horizons through the circles of kin selection, reciprocal altruism and group selection. Although I adopt his expressive metaphor, I differ from him in seeing the advance of human moral horizons in terms of the evolution of progressively more elaborate *equilibrium selection devices.* In spite of their importance, neither reciprocal altruism nor kin selection are concerned with how equilibria are *selected* in the game of life. Reciprocal altruism is about how equilibria in the game of life are *sustained.* Kin selection is about the manner in which *payoffs* should properly be calculated in family games. Reciprocity and kinship therefore form part of the ground on which the expanding circles of an evolving morality need to be drawn. My circles therefore all lie in the domain of *group selection* – a subject on which Singer (1980) is more than a little tentative because of a major controversy in biology over the use of this term. It is therefore necessary for me to stipulate that I do not intend group selection

2. The exceptions mostly concern sex and the family. One ought also to include relationships within small close-knit groups that seemingly trigger the same biological mechanisms that evolution has provided for use within the family.
3. The first volume *Playing Fair* is published at the time of writing (Binmore 1994). The second volume *Just Playing* is in preparation.

to be understood in the discredited sense attributed to Wynne-Edwards (1962) by Williams (1974) and others. Their criticisms do not apply to my use of the term because I restrict attention to groups whose organization cannot be destabilized by deviant insiders.

Reciprocity

Before turning to the expanding circles of group selection, it is necessary to say something about the ground on which such circles are drawn – the sets of equilibria in the game of life from which a selection must be made.

Axelrod (1982) is largely responsible for popularizing the fact that it is possible for highly cooperative outcomes to be sustained by equilibrium strategies using I'll-scratch-your-back-if-you'll-scratch-mine principles in indefinitely repeated games. Each player continues to cooperate so long as his fellows reciprocate, but plans to switch to a punishment strategy should anyone deviate. The simplest such strategy is perhaps TIT-FOR-TAT in the indefinitely repeated Prisoners' Dilemma. However, the folk theorem of repeated game theory shows that TIT-FOR-TAT is just one of an enormous number of strategies that can be used to sustain cooperation among players whose motivations are entirely selfish.

But what have the equilibria of repeated games played by selfish agents to do with Trivers' (1971, 1985) notion of *reciprocal altruism*? Altruism should surely involve some element of self-sacrifice. But no self-sacrifice is involved when Adam and Eve both play TIT-FOR-TAT in the indefinitely repeated Prisoners' Dilemma. They continue to cooperate by playing *dove* because they know that the victim of an exploitation attempt will retaliate by playing *hawk* until the deviant shows his contrition by cooperating again. But when the cooperative arrangement is working well, the darker side of their relationship will remain invisible to an observer. He will simply see Adam and Eve cooperating at every opportunity. A kibitzer may therefore be tempted to attribute their behavior to altruism. We are much moved, for example, by the mutual affection exhibited by pairs of lovebirds, but they arguably stick close to each other only because their partner is likely to be unfaithful if not watched continuously.

Reciprocal altruism is therefore something of a misnomer. One should rather say that the reciprocity mechanism makes some of the benefits of altruism available without the need for anyone to love his neighbor. In brief, Mr. Hyde is just as capable of getting on with his fellow men in repeated situations as Dr. Jekyll.

The possibility of reciprocal altruism assures us that Adam and Eve have a wide variety of cooperative outcomes that can be sustained as equilibria in an indefinitely repeated game of life. However, the question of how feasible outcomes are *sustained* is downplayed in this paper so that attention can be concentrated on how outcomes are *selected* from the set of all feasible possibilities. It seems to me that the origins of the equilibrium selection devices that we discuss under the heading of fairness or justice must be sought by looking first at the manner in which animals related by blood play the game of life that Nature deals to them.

Kinship

Hamilton's (1963, 1964) notion of *kin selection* concerns relations within the family. People related by blood share genes. A gene that modifies some piece of behavior will therefore be replicated more often if it takes into account, not only the extra reproductive opportunities that the modified behavior confers on a host who carries the gene, but also the extra reproductive opportunities it confers on those amongst the host's relatives

who also carry the same gene. The point was famously made in a semi-serious joke of J.B.S. Haldane. When asked whether he would give his life for another, he replied that the sacrifice would only be worthwhile if it saved two brothers or eight cousins!

Books on behavioral ecology (behavioral ecology) like those of Barash (1982), Hughes (1988), Krebs and Davies (1984) or Ridley (1993) are ultimately convincing because biologists are able to appeal to case studies in their thousands. Cooperative behavior in animals attributable to kin selection is very widespread indeed. African hunting dogs, for example, regurgitate food to help out a hungry pack brother. The evidence is particularly telling when special circumstances are revealed that explain why an apparent counterexample to the basic theory is in fact just one more supportive case. The explanation for the spectacular level of social organization amongst *hymenoptera* like bees and ants is particularly compelling. However, fascinating though it is, I plan to make no attempt to offer any kind of introduction to this huge literature. In particular, as everywhere else in the paper, the games of sibling rivalry to be discussed are not intended to be realistic. Their purpose is simply to indicate why it would be a mistake to proceed on the assumption that blood relatives will act as though selfishly seeking to maximize their own reproductive potential at the expense of the prospects of other members of the family.

Little birds in their nest agree

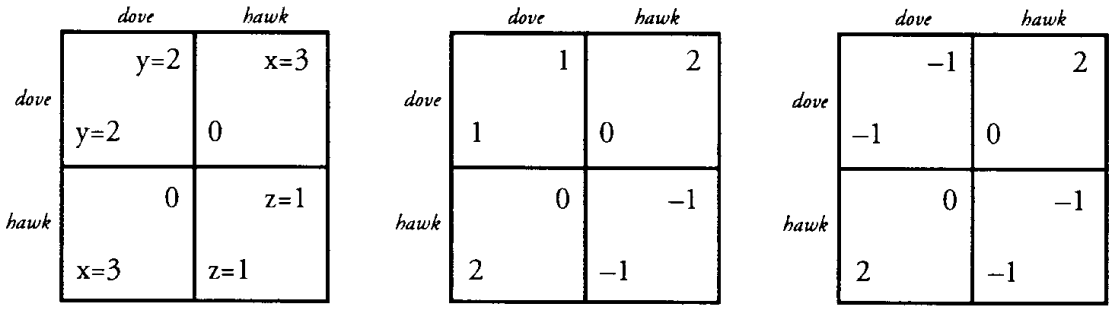
Haldane's aphorism was based on the fact that we share half our genes with a brother and one eighth of our genes with a cousin. A gene that programs us to save a cousin therefore has one eighth of a chance of saving a copy of itself. Hamilton's (1964) famous rule is an attempt to quantify the extent to which such considerations should be expected to result in one player making sacrifices on behalf of a relative. However, we shall have to wait some time before encountering this rule, and even then it will not be quoted in the standard form,[4] since it seems more useful for the purposes of this article to follow Bergstrom (1995) in stating the results in the more abstract language of game theory.

In order to make a start on the kinship problem, imagine that children are always born in pairs. Two siblings, Adam and Eve, will then occupy the same nest in their infancy. In the nest they play a sibling rivalry game, the outcome of which affects their eventual ability to pass their genes to the next generation. To keep things simple, it will be assumed that the sibling rivalry game is symmetric, with only two pure strategies *dove* and *hawk*. Some possible candidates for the sibling rivalry game are the Prisoners' Dilemma, Chicken, the Stag-Hunt Game of Figure 1.

The payoff $\pi(x,y)$ to Adam if he uses pure strategy x and Eve uses pure strategy y is to be interpreted as his biological fitness – the average number of extra offspring above some fixed background level π that he will produce as a result of the strategy pair (x,y) being played in the game.[5] Since the sibling rivalry game is symmetric, Eve's fitness is then $\pi(y,x)$.

When considering the evolutionary stability of a population, it is important bear in mind that the fact that Adam has fitness

4. Which says that altruistic behavior should be anticipated when $Br>c$, where B is the benefit to the recipient of the altruistic act, C is the cost of the act, and r is the degree of relationship between the benefactor and the beneficiary (who is assumed to be unique).
5. The fact that the number of children in a family is fixed at two does not contradict this interpretation of $\pi(x,y)$. Adam may raise several families in the breeding season should he survive so long.



(a) Prisoners' Dilemma (b) Chicken (c) Battle of sexes

Figure 1: Adam is the row player. His payoffs lie in the SW position of each cell. Eve is the column player. Her payoffs lie in the NE position of each cell.

$\pi(y,x)$ and Eve has fitness $\pi(x,y)$ is a secondary consideration. The central issue is the rate at which the genes they carry are replicated. Matters are simplest in the case of a unisex species in which all siblings have the same genes as their mother and hence are clones. If the behavior of children in the nest is entirely genetically determined, all the children of the same mother will necessarily choose the same strategy. If normal families choose *dove* and mutant families choose *hawk*, the rates at which genes are replicated in normal and mutant families increase by $\pi(dove,dove)$ and $\pi(hawk,hawk)$ respectively. If the former exceeds the latter, a population of normals will be evolutionarily stable. Any bridgehead of mutants will then eventually disappear because mutant genes replicate more slowly than normal genes.

It is, of course, no accident that this reasoning parallels the argument behind the Paradox of the Twins (Binmore, 1994). After all, Adam and Eve are indeed twins in this context. Evolution should therefore be expected to generate cooperation in the one-shot Prisoners' Dilemma, even though such behavior is not in equilibrium when Adam and Eve choose independently.

In more general games, one can summarize the behavior that a successful gene will instil in a population of clones using a crude version of Kant's categorical imperative: Maximize your payoff on the assumption that your siblings will choose the same strategy as yourself. In game-theoretic terms, the gene that ends up controlling the population will program Adam and Eve to optimize in the *one-player* game whose strategies are the same as in Adam and Eve's sibling rivalry game, but in which the payoff that results from choosing x is $\pi(x,x)$. It would therefore be a mistake to use the sibling rivalry game as the basis of a game-theoretic analysis of the behavior to be expected from Adam and Eve. Identical twins whose behavior is genetically determined are not independent players at all. The only true player is the single gene package that pulls their strings.

A more delicate analysis is necessary when studying sexual species. In the simple haploid case, the behavioral trait to be studied is determined by a single gene inherited with equal probability from the mother or the father. The population size will be assumed to be very large, and all individuals who survive to breeding age will be assumed to pair at random. To test a population for evolutionary stability, assume that a mutant gene has taken control of a very small fraction ε of a population that is currently pairing to raise families. The same fraction of their children will also be mutants, and so we must turn our atten-

tion to their grandchildren. The question is whether the fraction of mutants in this generation is greater or smaller than ε. If the latter, then the normal population is evolutionarily stable.

Since the number of mutant parents is small, a normal child will almost certainly come from a marriage in which both parents are normal, and a mutant child will almost certainly come from a mixed marriage with one normal parent and one mutant parent. As Bergstrom (1995) observes, the question of evolutionary stability therefore reduces to comparing the fitnesses of normal children from normal marriages with that of mutant children from mixed marriages.[6] Since the sibling of a mutant child in a mixed family has half a chance of being a fellow mutant and half a chance of being normal, a criterion for a normal population to be evolutionarily stable is therefore

$$\pi(\textit{dove, dove}) > \tfrac{1}{2}\pi(\textit{hawk, hawk}) + \tfrac{1}{2}\pi(\textit{hawk, dove}). \qquad (1)$$

Criteria of this kind seem to have been first obtained by Grafen (1978, 1984) and are used by behavioral ecologists to modify Hamilton's (1964) "inclusive fitness" to a new notion of "personal fitness" (Hines and Maynard Smith (1978:20)). However, both of these modifications to the classical notion of biological fitness are mentioned here only so that I can disclaim any intention of referring to them again in what follows. When I speak of a fitness, it will aways refer to the reproductive success of a particular individual.

Bergstrom (1995) refers to (1) as a semi-Kantian criterion, since the rule of behavior that a successful gene will instil in a haploid population is a hybrid of Kant's categorical imperative and the rule that was lightheartedly said in Section 2.4.3 of Binmore (1994) to be Nash's categorical imperative. The hybrid rule instructs you to maximize your payoff on the assumption that half the time your siblings will choose the same action as yourself and half the time they will not react to your behavior. Bergstrom (1995) describes the outcome as a symmetric, strict Nash equilibrium of the two-player game

6. To check the validity of criterion (1), we can begin by estimating the number N of grandchildren of all types and the number M of mutant grandchildren. If (1) is correct, it must be equivalent to the requirement that $M/N<\varepsilon$ for sufficiently small values of $\varepsilon>0$. To this end, let F be the number of families and let $\pi(x,y)=\pi+p(x,y)$ be the total number of children that an individual will have on average if he uses strategy x and his sibling uses strategy y.

The fraction of marriages in which both parents are normal is $(1-\varepsilon)^2$, the fraction in which one parent is normal and the other mutant is $2\varepsilon(1-\varepsilon)$, and the fraction in which both parents are mutants is ε^2. If we neglect terms of order ε or higher, it follows that we can proceed as though all marriages are normal in estimating the total number N of grandchildren of all types. Thus N is approximately $2Fp(\textit{dove, dove})$. In estimating M, we need to retain terms of order ε, but neglect all higher order terms. We can therefore proceed as though all mutant children come from mixed marriages, of which there are approximately $2\varepsilon F$. For the same reason, the number of mutant grandchildren – who will nearly always be born into a mixed marriage – is approximately the same as the number of mutant children. One quarter of the time, both children from a mixed marriage will be normal and so we have nothing to count. One quarter of the time, both children will be mutants. The number of mutant grandchildren deriving from such a family is $4\varepsilon Fp(\textit{hawk, hawk})$. One half of the time, one child will be normal and the other a mutant. Since only one of the children in such a family will be producing mutant progeny, the number of mutant grandchildren is then $2\varepsilon Fp(\textit{hawk, dove})$. The total number M of mutant grandchildren is therefore approximately

$$\tfrac{1}{4}\,4\varepsilon Fp(\textit{hawk, hawk}) + \tfrac{1}{2}\,2\varepsilon Fp(\textit{hawk, dove}).$$

It follows that $M/N<\varepsilon$ for small values of $\varepsilon>0$ if and only if (1) holds.

whose strategies are the same as in Adam and Eve's sibling rivalry game, but in which the payoff that Adam gets when he chooses x and Eve chooses y is

$$V(x,y) = \tfrac{1}{2}\pi(x,x) + \tfrac{1}{2}\pi(x,y). \qquad (2)$$

However, I think one does better to conceive of the situation as a leader-follower game[7] in which the leader first chooses y with the aim of maximizing $\pi(x,x)$ and the follower then responds by choosing x with the aim of maximizing $\frac{1}{2}\pi(x,x) + \frac{1}{2}\pi(x,y)$. After the choice of y, the follower is then involved in a *one-player* game.

The Kantian nature of the intrafamilial ethics generated by considerations of this kind is apparent when the sibling rivalry game is the one-shot Prisoners' Dilemma. If Adam and Eve were really playing this game in the nest, then they would both choose *hawk*, since this strongly dominates *dove*. But the real players are the genes that control their behavior and, as we have seen, they have a different game to play. When Adam and Eve are clones, the single gene package that controls them will program them to cooperate by playing *dove*, for the reasons that Kant gave to motivate his categorical imperative. When they are siblings in a haploid population, they will be programmed to employ the semi-Kantian rule. If $2y>x+z$ and $y>z$ in the general one-shot Prisoner's Dilemma of Figure 1(a), the result will be that both cooperate by playing *dove*. They will similarly play like doves when the sibling rivalry game is Chicken or the Stag-Hunt Game.

Of course, the situation analyzed above has been absurdly oversimplified. Sibling rivalry games have an indefinite number of players and need not be symmetric. Nor is the degree of relationship between the players likely to be sharply defined, if only because of the possible presence of cuckoos. Animals do not usually mate at random, nor are populations effectively infinite. Even the notion of a gene as an atomic entity becomes suspect when the molecular realities are closely studied. Above all, the manner in which genes control behavior must surely be much more complicated than the manner in which they control eye color. The human species is not even haploid.[8] However, we need to turn next to an even more fundamental difficulty that kicks in when the human species is under discussion.

Learned or instinctive behavior?

The preceding discussion was set against a background of baby birds sharing a nest in order to emphasize that Adam and Eve's behavior was assumed to be entirely under the control of their genes. My reading of the literature suggests that sociobiologists sometimes forget their fundamental commitment to coevolution when reacting against the *tabula rasa* theory popular among social scientists by taking the same assumption for granted when extrapolating sociobiological conclusions to humans. They then treat *homo sapiens* as

7. To defend the leader stage of the game, one needs to appeal to some sort of group selection argument of the type discussed later on.
8. We are a diploid species which carries *two* genes at each locus. Even in a single locus model, it is therefore necessary to take account of two genes. If the mutant gene is dominant, the necessary considerations are essentially the same as in the haploid case, but matters become more complicated when the mutant gene is recessive. Bergstrom (1995) shows that $V(x,y)$ must then be replaced by $W(x,y) = \frac{1}{5}\pi(x,x) + \frac{3}{5}\pi(x,y) + \frac{1}{5}\pi(y,x)$. Bees and ants are haplodiploid – unfertilized eggs produce haploid males and fertilized eggs produce diploid females, which goes a long way towards explaining why such species can reach such high levels of sociality (Hamilton, 1964, Alexander *et al*, 1991).

homo behavioralis – a stimulus-response machine programmed directly with behavior like a chocolate-dispensing machine. The temptation is to follow Kant (1785:63) in thinking that any mundane purpose that humans might exist to pursue, such as ensuring that the genes that pull their strings are passed on to the next generation, would be best served by hominids who do not reason at all, but are simply hardwired with the optimal response to each relevant stimulus.[9]

But whatever sociobiologists may or may not believe about the extent to which human behavior is instinctive, my own view is that the game of life is too complicated for it to be possible for us to be hardwired with optimal strategies for all its subsidiary games. For most purposes, our genes therefore do not program us directly with behavior like *homo behavioralis.* It seems more realistic to proceed on the assumption taken for granted by folk psychology that our cognitive processes really do involve some use of the preferences and beliefs of *homo economicus* that revealed preference theory treats as convenient fictions.

Many of our personal preferences are doubtless genetically determined, like hunger, thirst and the sexual urge. Perhaps some of our beliefs are also hardwired – making the world not only more strange than we imagine, but more strange than we can imagine. But some preferences and most beliefs must surely be acquired. That is to say, our genes do not always insist that we prefer or believe specific things; in some contexts they insist only that we organize our cognitive processes in terms of preferences and beliefs. On this view, we come equipped with algorithms that not only interpret the behavior patterns that we observe in ourselves and others in terms of preference-belief models, but actively build such models into our own operating systems. As Hume (1789:420) remarks, "Nothing has a greater effect both to increase and diminish our passions, to convert pleasure into pain, and pain into pleasure, than custom and repetition."

It seems likely that the psychological mechanisms involved in learning new behavior or acquiring new preferences or beliefs are many and varied. However, I plan only to speculate about the minimal set of psychological mechanisms that seem consistent with the story I have to tell about human kin selection.

It is probably uncontroversial to suggest that we are natural imitators. Like proverbial monkeys, we tend to copy what we see others doing, whether the behavior makes much sense or not. But neither humans nor monkeys are totally uncritical. We test our newly acquired behaviors against our preferences, as expressed through our emotional responses. In short, we ask ourselves whether we like the consequences of our new behavior. If the behavior is found wanting, we seek to refine it, or else return to a tried-and-true alternative.

But where do we get the preferences used for this purpose when these are not wired in at birth? My guess is that we come equipped with algorithms that operate *as though* they were employing the principles of revealed

9. Kant (1785:64) thought that reason was therefore superfluous to our mundane needs and hence must exist to serve the transcendental purpose of creating a "will which is good". Such a will, so he argued, would necessarily operate his categorical imperative. As usual, trying to follow Kant's thoughts is like entering a Looking-Glass World. For example, we have just seen that a defense can be mounted for the type of intuition which led him to the categorical imperative in the case when the players belong to the species *homo behavioralis.* But this is precisely the case his own argument excludes!

prefence theory to deduce preference-belief models from consistent sets of behavior. The process being proposed is recursive rather than circular. A well-established set of behaviors employed in a particular context is first encoded as preference-belief model. This preference-belief model is then used to test or refine new behaviors[10] before they are admitted into an individual's repetoire of habituated responses. An adjustment period follows in which behaviors are refined until full consistency is achieved, whereupon a revised preference-belief model is constructed.

The evolutionary advantage of such an inductive process is that new behaviors are tested against past experience in an internal laboratory before being put to use in the gladiatorial arena of life. If the environment is sufficiently stable that our past experience is relevant to present challenges, it then becomes possible to assimilate new behaviors quickly without taking large risks. Since a successful innovation by one individual can swiftly spread through a whole culture by imitation, the mechanism therefore makes us an unusually flexible animal.

But a price has to be paid for our flexibility. The fact that we must learn how to behave makes us a second-best species, in that Nature loses fine control over the way we play most games. In particular, human identical twins cannot emulate the behaviorally hardwired species for which the Paradox of the Twins can be made to work. If the preferences which mediate human behavior adequately reflect success or failure in the evolutionary race, then I shall shortly be arguing that we are condemned to choose *best replies* in the games we play – or else be displaced by rivals who do. But if everyone chooses a best reply, then the result will be a Nash equilibrium of the game.

Since our nature forces us to play Nash equilibria in games, our species must live with the cost of being unable to sustain first-best outcomes in games like the one-shot Prisoners' Dilemma. On the other hand, we are able to avoid the fate of a behaviorally hardwired species when playing games to which its members are not adapted. Rather than playing whatever third-best strategy might happen to be triggered by the available stimuli, humans enjoy the benefit of having the potential to learn a second-best strategy in any game whatsoever, whether it figured large in our evolutionary history or not.

Although a man who operates on the psychological principles I have been proposing will eventually respond to a new challenge as though he were *homo economicus*, it is important to insist yet again that it does not follow that he is more than dimly aware of what is going on in his head – even when the stories he tells himself and others about his inner life are entirely consistent with his observed behavior.

Introspection seems particularly problematic when trying to assess the extent to which we sympathize with the fate of our relatives. How much do I love my brother? Is love even the right word to describe my feelings? Personally, I find myself unable to give adequate answers to such questions – and I suspect that the more definite views offered by others owe more to romantic fiction than genuine self-knowledge. In accordance with the model I have been describing, it seems to me that we unconsciously learn how we feel towards our relatives by experimenting with different behaviors and observing the emo-

10. Such new behaviors may be constructed *de novo* from the preference-belief model, but calculating behavior of this kind must surely be comparatively rare compared with behavior acquired through imitation or serendipity.

tional responses of our bodies. At the end of the process, I may have no better way of articulating what I have learned to feel than saying that I love my brother so much that I am even prepared to lend him my blue suede shoes to impress his date tonight. But the fact that introspection seldom allows us to quantify our emotions adequately in a more direct way does not imply that they do not control our behavior very closely. On this subject at least, folk psychology surely hits the nail right on the head.

Hamilton's rule

Substantive consequences follow from using a model of man whose genes control him by manipulating his preferences rather than his behavior. In particular, when discussing humans, it becomes doubtful whether it is such a good idea to replace Hamilton's rule by one or other of the neo-Kantian criteria proposed as corrections to Hamilton's rule by Grafen (1979) and Bergstrom (1995). In fact, I shall now argue that Haldane got things essentially right when he proposed that a human should count a brother's fitness as being half as valuable as his own.

The first step is to justify the claim made above that humans are condemned to play Nash equilibria in the game of life. I think that this is most easily seen by comparing what is involved when a human deviates from a socially determined norm with what happens when a mutant baby bird deviates from a genetically determined norm.

Imagine that Adam and Eve are equipped with preferences that are either genetically determined or else have been distilled from habituated behavior acquired in the past. In either case, it is important that the preferences remain relatively stable as new behavior is assimilated through a combination of individual trial-and-error adjustment and social imitation. I follow Aumann (1987b) in seeing the essence of such interactive learning as a two-stage process in which we first receive a social signal that tells us how to behave, and then test the behavior against our preferences to see whether we wish to follow its recommendation. Such considerations lead Aumann to the notion of a correlated equilibrium, but matters are simpler here, because a player in a prehistoric family game is unlikely to have been able to receive signals from society of which his relatives were unaware. Operating a correlated equilibrium under such circumstances just reduces to specifying how a Nash equilibrium is to be selected. For example, in the Battle of the Sexes, the social signal might simply specify that the Nash equilibrium (*dove, hawk*) be played. But it could require that a coin be tossed, with the equilibrium (*dove, hawk*) being played if it falls heads and the equilibrium (*hawk, dove*) if it falls tails.

It is important that the social norm in use finally advocates the use of a Nash equilibrium, because the players are assumed to test whatever recommendation is made to them against their preferences. In practice, this means that they will occasionally experiment, either hypothetically or actually, with strategies that have not been recommended, in order to discover whether they can thereby gain a greater payoff. As discussed at great length in Binmore (1994, section 3.4.1), deviations by different players must then be expected to be *uncorrelated*. Even if Adam and Eve are identical twins, the fact that Adam happens to try out a deviation from the social norm just before bedtime on Tuesday provides no reason for supposing that Eve will simultaneously select precisely the same moment to deviate. The condition for the ocial norm to be stable or self-policing is therefore that it recommend that each player make a best reply to the behavior it recommends to the other players. In other

words, the social norm must coordinate the players' behavior on a Nash equilibrium. This situation contrasts sharply with the case of behaviorally hardwired identical twins. The presence of a mutant gene in one player then guarantees its presence in the other. A deviation from the norm induced by the mutant gene in one player will then be matched exactly by a precisely similar deviation in the other.

Given the model proposed above for human behavior, what preferences will Nature write into the game of life played by relatives whose degree of relationship is r? If Adam and Eve are identical twins, $r=1$ because they then share all their genes. If they are nonidentical twins, $r=\frac{1}{2}$ because they then share half their genes. If they are supposedly brother and sister, then r is something less than a half because of the risk of some unfaithfulness on the part of their mother. If they are known to have the same mother but different fathers, then $r=\frac{1}{4}$. If they are known to be full cousins, then $r=\frac{1}{8}$. If they are members of a wider kingroup, then r is some smaller but positive number.

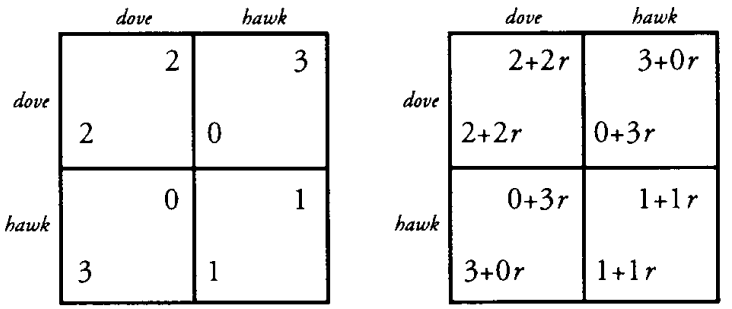
In all these cases, it seems to me almost tautological that evolution will eventually program Adam and Eve with personal utility functions u_A and u_E that are computed from their respective fitnesses using Hamilton's rule in the form:

$$u_A(x,y) = \pi(x,y) + r\pi(y,x), \qquad (3)$$
$$u_E(x,y) = \pi(y,x) + r\pi(x,y).$$

We will then be dealing with a case in which an individual's personal preferences are not narrowly selfish. When Adam and Eve are relatives, they will *sympathize* with each other's reproductive aspirations. Each player explicitly includes his relative's biological fitness as an argument in his personal utility function, which then determine his payoffs in the game of life. As a result, one is likely to see siblings cooperating even when their sibling rivalry game is the one-shot Prisoners' Dilemma. One can argue that they are nevertheless not behaving altruistically, since each is in fact optimizing in the game of life that they actually are playing. However, we shall get no closer to true altruism than in the kin-selection examples of this section.

The reasoning that leads to (3) is much simpler than that which led to the semi-Kantian rule embodied in (2). In the case of baby birds in their nest, the very fact that a mutant gene is planning to cause a player to deviate provides the gene with information. The gene learns that there is a significant probability that the other player will deviate simultaneously. Equally importantly, it knows that if the other player does match the deviation, then both players will be carrying the mutant gene and so its payoff will be doubled. On the other hand, if the other player doesn't match the deviation, then he isn't carrying the mutant gene and hence his fitness does not contribute to the gene's payoff.

Such complexities do not apply in the human case – nor presumably, to primates and other animals that transmit their culture from one generation to the next. After Adam deviates, a gene that modifies his preferences knows no more about its presence in Eve than it did before. If the degree of relationship between Adam and Eve in a human family is r, then one of Adam's genes will continue to believe that it is present in Eve with probability r even after Adam has deviated from the social norm. Assuming that Adam is genetically programmed to maximize something, the propogation rate of a controlling gene will therefore be optimized if the formula used in calculating whatever is to be maximized is (3). Similarly, given that the propogation rate of Adam's genes is optimized by maximizing some particular function, a gene that controls Adam's learning algorithm will do best if the



(a) Sibling rivalry game (b) Game of life

Figure 2: Sibling rivalry in the Prisoners' Dilemma.

algorithm eventually teaches him to maximize that function.

As usual, the benchmark question is how things go when the sibling rivalry game is the one-shot Prisoners' Dilemma. The game of life obtained by transforming the Prisoners' Dilemma of Figure 2(a) using Hamilton's rule in the form (3) appears in Figure 2(b). In the transformed game, *dove* dominates *hawk* when $r \geq \frac{1}{2}$. It is therefore optimal for the players to cooperate.

The units in which biological fitnesses are expressed are expected numbers of offspring. However, nothing says that Adam and Eve must value each other's children equally. Indeed, Hamilton's rule requires that each player value his sibling's children at half the rate at which he values his own. When they are siblings, Adam therefore regards Eve's units of fitness as being worth half his own units of fitness. Similarly, she regards his units of fitness as being half as valuable as hers. One might then say that they act as semi-utilitarians. Just as a semi-Kantian chooses x to maximize $V(x,y) = \frac{1}{2}\,\pi(x,x) + \frac{1}{2}\,\pi(x,y)$, so a semi-utilitarian chooses x to maximize $V(x,y) = \frac{1}{2}W_H(x,y) + \frac{1}{2}\pi(x,y)$, where $W_H(x,y) = \pi(x,y) + \pi(y,x)$ is the classical utilitarian social welfare function.

Adam and Eve share a common standard for making interpersonal comparisons of their fitnesses only in the extreme case when they are identical twins, so that $r=1$. They then become classical utilitarians, whose personal payoffs in their game of life are found simply by adding their biological fitnesses. The game of life then becomes a common interest game. In a common interest game, Adam and Eve receive equal payoffs whatever the outcome of the game may be. No scope for conflict therefore exists. Their joint interest lies in maximizing the sum of their fitnesses, even if this means that Adam must lay down his life for Eve. However, unlike Kantian identical twins, whose behavior is genetically determined, a human pair of identical twins continue to play a *two-person* game of life. They therefore retain the capacity to screw things up by coordinating on a Pareto-inferior equilibrium – leaving their family line vulnerable to elimination by group selection as families who succeed in coordinating on a Pareto-superior equilibrium forge ahead in the reproduction race.

Group selection

The mention of *group selection* is likely to generate knee-jerk hostility because the term is associated with the discredited theory that evolution favors mutations which benefit the whole species in which they occur. This theory was exploded long ago by Williams

(1974), Dawkins (1976) and numerous others. Indeed, one only has to read the title of Dawkin's *Selfish Gene* to realize that it is obvious that evolution must either operate at the level of the gene or not at all, because genes are what actually get replicated. As long as a mutant gene replicates faster than a normal gene, it will take over whether or not the species as a whole benefits.

But there are other types of group selection that are regarded as being entirely respectable in the behavioral ecology literature (Boyd and Richerson 1990, Grafen 1984, Nunney 1985, Wade 1985, Wilson 1983, Wilson and Sober 1988.) I plan to restrict the term in this paper to a particularly narrow sense that excludes group organizations that are not internally stable, and therefore would not survive even if they were not competing with other groups. Such a usage has the advantage that "selfish gene" criticisms of the notion are obviously not relevant.

In game-theoretic terms, a necessary condition for a group to be stable against deviations from within is that it operate an equilibrium of whatever game of life its members play with each other. However, games typically have many equilibria. Some of these equilibria will be better than others, in the sense that animals from a group using one equilibrium may do better on average than animals from a group using another equilibrium. This will certainly be the case if the equilibrium used by the first group is a Pareto-improvement on the equilibrium used by the second. Groups using a Pareto-superior equilibrium will therefore grow in size or number at the expense of groups using a Pareto-inferior equilibrium. Eventually, the inferior groups will disappear. Evolution will then have operated to favor one kind of group organization over another.

Since a stable group organization corresponds to an equilibrium in the game of life its members play, group selection is an equilibrium selection mechanism that tends to eliminate equilibria that have Pareto-superior rivals. How long the mechanism takes to eliminate a Pareto-inferior equilibrium currently being operated by all groups will depend on how likely it is that any single group will spontaneously shift to a Pareto-superior equilibrium. If mutation is the only source of relevant variability, the expected waiting times may well exceed the expected lifetime of the species. For example, in the order *hymenoptera* that includes ants, bees and wasps, we find different species operating many levels of social contract, from the Hobbesian "war of all against all", in which solitary individuals seek to maximize their own fitness with no regard to the help that they may give or receive from others, to the Rousseavian ideal, in which all individuals in a colony seemingly subordinate their own self-interest to a "general will". Evolutionary biology therefore provides no guarantee that a species will learn to cooperate even when the conditions are seemingly favorable.

Group selection in humans

Insect species have to wait for chance to shift a society from one equilibrium to a more cooperative alternative. Societies operating the less efficient equilibrium will then eventually disappear if they are competing for the same resources. But human societies are more resilient, since our capacity to imitate makes it possible for one society to borrow the cultural innovations of another. However, my guess is that Nature has made us even more flexible. I think it likely that we are genetically programmed with algorithms that help to immunize our societies against competition from innovative rivals. Such algorithms actively seek out Pareto-improving equilibria as these become available through changes in the environment in which we live.

My guess is that the fairness norms which seem universal in human societies have evolved primarily for this purpose. Their principal role is to single out one of the many equilibria that will typically be available as Pareto-improvements on the *status quo* without the necessity for damaging and potentially destabilizing internal conflict.

Binmore (1994) explains at length why social tools of this type, that we successfully employ in deciding what is fair in small-scale situations, have the potential to be harnessed for use in finding answers to large-scale problems for which our evolutionary history has provided us with no solution techniques. In terms of the expanding circle metaphor to be pursued below, this is to suggest that we consciously seek to increase the radius of the domain in which we operate as moral animals. However, such a proposal has no hope of working unless we are realistic about the second-best nature of the moral systems that we currently operate. This is why I am so anxious to trace the natural origins of our intuitions about justice rather than attributing them to inspiration from the realms of metaphysics.

Expanding the circle

To summarize the story so far, the claim is that the evolution of human ethics began as a consequence of group selection operating in family games. It began in the family because the equilibrium selection problem for games played by close relatives was easier for Nature to solve than in games played by strangers. The reason is that humans actively sympathize with their relatives by building a direct concern for their welfare into their personal utility functions. In extreme cases, we have seen that this effect may be enough to convert a sibling-rivalry game that looks like the one-shot Prisoners' Dilemma into a game of life in which the only equilibrium is Pareto-efficient.

The more likely someone we encounter frequently is to be a close relative, the more we tend to sympathize with their needs and concerns. When we get to distant kinfolk whom we seldom encounter, the degree to which we sympathize with them becomes small. But my guess is that we still retain more than enough capacity to sympathize even with absolute strangers to explain the "warm glow" feelings that lead us to leave tips in restaurants that we never expect to visit again or to make donations to charitable causes that are small compared with our income.

Such vestigial warm-glow feelings provide an inadequate foundation on which to build a modern state. It is simply not in our nature to love strangers in the same way that we love our near and dear. This is not to deny the existence of rare saintly individuals. Nor that we are all capable of acts of great self-sacrifice on rare occasions. But a utopian state that relies on saintly behavior from its citizens on a day-by-day basis will just not work. When moral behavior expanded from the extended family to the world at large, it did so in a more subtle way than by training us to love all men as we love our brothers. But to understand the mechanism, it is necessary to give up the idea that Nature might have changed the structure of the game we play with strangers in the same way that it changed the structure of sibling rivalry games. We have to focus attention instead on how evolution succeeded in shifting us from one equilibrium to another in a *fixed* game.

My guess is that the moral circle sometimes expands through players misreading signals from their environment and so mistakenly applying a piece of behavior or a way of thinking that has evolved for use within some inner circle to a larger set of people or to a new game. When such a mistake is made, the players attempt to play their part in

sustaining an equilibrium in the game played by the inner circle without appreciating that the game played in the outer circle has different rules. For example, Adam might mistakenly treat Eve as a sibling even though she is a complete stranger. Or he might mistake a one-shot game for an indefinitely repeated game.

A strategy profile that is an equilibrium for an inner-circle game will not usually be an equilibrium for an outer-circle game. The use of the inner-circle equilibrium strategy in the outer-circle game will therefore usually be selected against. But playing the outer-circle game as though it were the inner-circle game will sometimes result in the players coordinating on an equilibrium of the outer-circle game. The group will then have stumbled upon an *equilibrium selection device* for the outer-circle game. They succeed in coordinating on an equilibrium in this game by behaving as though they were playing another game with a more restrictive set of rules.

In summary, we need to turn our attention away from circles within which Adam and Eve sympathize with each other's plight to wider circles in which the extent of their sympathetic identification is too weak to be significant. Insofar as they can, *empathetic preferences* then have to substitute for the sympathetic preferences that operate within families. Such empathetic preferences have been extensively studied by Harsanyi (1977) and others under the name of extended sympathy preferences. Such a preference is expressed when someone says that he would rather be Adam drinking a cup of coffee than Eve drinking a cup of tea.

It is important to recognize that the same internal algorithms that allow us to use sympathetic preferences also allow us to handle empathetic preferences. When one gets down to brass tacks, both sympathetic and empathetic utility functions are computed simply as a weighted sum of the payoffs in a game. In the case of sympathetic preferences, the game is some analogue of a sibling rivalry game. In the case of empathetic preferences, the game is some analogue of the type of food-sharing game to be discussed shortly. Nature therefore did not need to invent some entirely new mechanism to bridge the gap between altruism within the family and fairness between strangers. She merely needed to supply some ramshackle scaffolding while adapting mechanisms that evolved to meet one set of needs to new and different purposes.

Empathy and fairness

Alexander (1974) and Humphrey (1976) are credited with the idea that we have large brains as a result of an arm's race within our species aimed at building bigger and better internal computing machines for the purpose of outwitting each other. This seems a very plausible speculation to me, but then the idea is one that would naturally attract a game theorist. But whether or not our capacity to empathize with our fellow men is the primary reason that we have bigger brains than other primates, it seems uncontroversial that we are genetically equipped to put ourselves in the position of others to see things from their point of view.

The importance of empathetic identification in helping us to predict the behavior of others will be obvious. Its role in facilitating learning by imitation may not be quite so apparent. Since the spread of social norms through imitation and education is an important backdrop to what follows, it may therefore be worthwhile to observe that Adam needs to understand *why* Eve behaved in a certain way if he is to know *when* the time has come to copy her behavior. To use his capacity for empathy to understand what triggered her behavior, he needs to imagine himself in her shoes with her preferences and her be-

liefs. It is then not such a long step to comparing his personal preferences with her personal preferences, just as he compares his fitness with his sister's fitness in a sibling rivalry game. However, before embarking on a discussion of such empathetic preferences, it is first necessary to say something about the type of equilibrium selection problem that they presumably first evolved to solve.

Insurance contracts and the original position

The device of the original position has been advanced by Harsanyi (1977), Rawls (1972) and others as a means of evaluating the fairness of political constitutions. Its use by Adam and Eve requires that they imagine that they have passed behind a veil of ignorance that conceals their current role in society. Behind this veil of ignorance, they bargain over the type of society that they will institute on leaving the original position. The fact that neither knows which role they will then occupy requires that each player must evaluate the available prospects using empathetic preferences.

Aside from discussing the modeling problems raised by the original position in detail, Binmore (1994) argues that its intuitive appeal lies in the fact that we *already* use similar fairness norms on a daily basis to find fair solutions of picayune bargaining problems – like who should wash how many dishes. But how might a fairness norm incorporating the principles of the original position have evolved? I look for the answer in the primitive food-sharing arrangements that anthropologists sometimes suggest mark the beginnings of human sociality.

If player I is lucky enough to have an excess of food this week, then it makes sense for him to share with player II in the expectation that she will be similarly generous when she is lucky in the future. Things are similar in the case of the alliances that operate within chimpanzee societies. One chimp come to the aid of an ally who is unlucky enough to incur the emnity of a powerful foe in the expectation that the service will be reciprocated when their roles are reversed.

If the players are relatives, such relationships will be easier to get off the ground, since each player will sympathize with the other to some degree. But the reciprocal arrangements built into such mutual insurance pacts can work perfectly well without any need to attribute altruistic motives to the players. Indeed, the folk theorem of repeated game theory tells us that we must expect there to be an embarrassingly large number of alternative equilibria amongst which a choice must be made. In deciding which equilibrium to operate, the players are therefore confronted with a classic bargaining problem.

The possible agreements are sharing rules that must be negotiated before the players know who will be lucky. In predicting how much each is likely to get from any particular rule, the players will use their common experience of how things have gone in the past to assign a probability p to the event that it will be player I who is lucky. My guess is that such mutual insurance pacts are more likely to have operated successfully between players of roughly equal prowess, but it is not essential for the argument that we take $p = \frac{1}{2}$. Whatever the value of p, each possible sharing rule determines an expected payoff to each of players I and II. The set T of all such payoff pairs is the set of feasible agreements for their bargaining problem. The *status quo* τ is the payoff pair that results if they operate without an insurance pact.

I don't suppose anyone knows to what extent our primitive ancestors bargained like buyers and sellers in a modern bazaar. The tradition is doubtless very ancient, but we don't need to assume that the bargaining was formalized in this particular way – or even

that the proposals and counterproposals made by the players were verbalized at all. It would come to much the same thing if the players simply acted out their proposals and counterproposals physically over a period of time when both were frequently being buffetted by the winds of fortune. But however the bargaining may have been done, the important point is that bargaining of some sort must have taken place under circumstances very close to those envisaged by Rawls (1972) and Harsanyi (1977) when they independently proposed the notion of the *original position.*

To see the similarity between bargaining over mutual insurance pacts and bargaining behind the veil of ignorance that separates individuals from their identities in the original position, think of players I and II as not knowing whether tomorrow will find them occupying the role of Mr. Lucky or Ms. Unlucky. It then becomes clear that a move to the device of the original position requires only that the players put themselves in the shoes of somebody else – either Adam or Eve – rather than in the shoes of one of their own possible future selves. However, on the face of it, a substantial gap still separates Rawls' or Harsanyi's proposed use of the original position to judge the fairness of political constitutions, and its use by our prehistoric ancestors in settling disputes over how a carcase should be divided. The same distinction separates Buchanan and Tullock's (1962) "veil of uncertainty" from Rawls' (1972) veil of ignorance. Dworkin (1981) similarly distinguishes between "brute luck" and "opportunity luck".

In a prehistoric insurance contract, the parties to the agreement do not have to *pretend* that they might end up either as Adam or as Eve. On the contrary, it is the reality of the prospect that they might end up as Mr. Lucky or Ms. Unlucky that motivates their writing a contract in the first place. But when the device of the original position is used to adjudicate fairness questions à la Rawls, then player I knows perfectly well that he is actually Adam and that it is physically impossible that he could become Eve. To use the device as recommended by Rawls and Harsanyi, he therefore has to indulge in a counterfactual act of imagination. He cannot become Eve, but he must pretend that he could. How is this gap between reality and pretence to be bridged without violating the Linnean dictum that Nature makes no jumps?

Natura non facit saltus

I see the step from the use of the device of the original position in negotiating mutual protection arrangements to its use in adjudicating fairness disputes as an example of how morality can expand from one circle to another. To reiterate the theory, people take a technique used within one circle of social problems and unthinkingly apply it to a wider domain of problems. In so doing, they continue to play by the rules of the game for which the technique originally evolved, not noticing – or pretending not to notice – that the rules of the game played in the wider circle may be quite different. Usually the result will not even be an equilibrium in the wider game and evolution will briskly sweep the experiment away. But sometimes the procedure will succeed in coordinating behavior on a Pareto-improving equilibrium of the wider game, whereupon group selection will move the evolutionary ratchet one further notch forward.

I have argued that the device of the original position has its roots in the need of early mankind to negotiate Pareto-improving insurance contracts. Earlier, I argued that the origins of empathetic preferences are to be found in the games played by kinfolk. I now

propose to argue that evolution somehow found a way to combine these two developments to create the equilibrium selection device of which we are dimly aware when making appeals to fairness or justice. My guess is that at least some of the necessary internal plumbing that allows us to operate this equilibrium selection device is genetically fixed – and hence the universal attachment across cultures to the basic notions of fairness and justice. However, I think it likely that the empathetic preferences that serve as inputs to the justice algorithm are almost entirely socially determined – and hence the different outcomes that result from using the justice algorithm in different societies.

As observed above, the leap from the use of the original position in self-policing insurance contracts to its use as a fairness criterion may seem unduly large. In negotiating an insurance contract, to accept that I may be unlucky is a long way from contemplating the possibility that I might become another person in another body. But is the difference really so great? After all, there is a sense in which none of us are the same person when comfortable and well-fed as when tired and hungry. In different circumstances, we reveal different personalities and want different things. When planning ahead under uncertain conditions, it would therefore not be surprising if we estimated our payoffs using the same wiring that we use when estimating payoffs in family games.

When planning ahead, a player computes his expected utility as a weighted average of the payoffs of all the people that he might turn out to be after the dice has ceased to roll. When choosing a strategy in a family game, a player takes his payoff to be a weighted average of the fitnesses of everybody in the family. In order to convert our ability to negotiate insurance contracts into a capacity for using fairness as a coordinating device in the game of life, all that is then needed is for us to hybridize these two processes by allowing a player to replace one of the future persons that a roll of the dice might reveal him to be by a person in another body who is to be treated in much the same way that he treats his sisters, his cousins or his aunts.

Conclusion

This paper reviewed Singer's (1980) approach to evolutionary ethics and expressed his three "expanding circles" of reciprocal altruism, kin selection, and group selection in game-theoretic terms. It endorsed Hamilton's rule in the case of humans and thereby found a possible evolutionary history for fairness norms that employ the principles of the device of the original position.

The use of such a fairness norm requires that Adam and Eve are equipped with empathetic preferences, but the origin of their empathetic preferences was left hanging in the air. Binmore (1994, 1996) argues that Adam and Eve's empathetic preferences should be seen as being determined by social evolution. In the medium run, they will then stabilize at an equilibrium of the underlying evolutionary process. At such an equilibrium, a common standard for making interpersonal comparisons of utility is established, and one can discuss the circumstances under which one is led to the utilitarianism favored by Harsanyi (1977) or the egalitarianism of Rawls (1972). However, these issues are beyond the scope of the current paper.

References

Alexander R. 1974. "The evolution of social behavior." *Annual Review of Ecology and Systematics*, 5:325-383.

– 1987. *The Biology of Moral Systems.* Hawthorne, New York.

Alexander R., K. Noonan and B. Crespi 1991. "The evolution of eusociality." In P. Sherman *et al*, editor, *The Biology of the Naked Mole Rat.* Princeton University Press, Princeton.

Aumann R. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55:1-18.

Axelrod R. 1982. *The Evolution of Cooperation.* Basic Books, New York.

Barash D. 1982. *Sociobiology and Behavior.* Elsevier, New York.

Barkow J., L. Cosmides and J. Tooby 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* Oxford University Press, Oxford.

Bergstrøm T. 1995. "On the evolution of altruistic ethical rules for siblings." *American Economic Review*, 85.

Binmore K. 1994. *Playing Fair: Game Theory and the Social Contract I.* MIT Press, Cambridge, Mass.
– 1996. *Just Playing: Game Theory and the Social Contract II.* MIT Press, Cambridge, Mass. (forthcoming)

Boyd R. and P. Richerson 1985. *Culture and the Evolutionary Process.* University of Chicago Press, Chicago.
– 1990. "Group selection among alternative evolutionary stable strategies." *Journal of Theoretical Biology*, 145:331-342.

Buchanan J. and G. Tullock 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy.* University of Michigan Press, Ann Arbor.

Chomsky N. 1959. A review of B.F. Skinner's "Verbal Behavior". *Language*, 35:26-58.
– 1965. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, Mass.
– 1980. *Rules and Representations.* Columbia University Press, New York.

Cosmides L. and J. Tooby 1994. Better than rational: Evolutionary psychology and the invisible hand. *Americal Economic Review (Papers and Proccedings)*, 84:327-332.

Cronin H. 1991. *The Ant and the Peacock.* Cambridge University Press, Cambridge.

Darwin C. 1871. *The Descent of Man and Selection in Relation to Sex.* Murray, London.

Dawkins R. 1976. *The Selfish Gene.* Oxford University Press, Oxford.

de Laguna T. 1965. "Stages of the discussion of evolutionary ethics." *Philosophical Review*, 15:583-598.

Dennett D. 1995. *Darwin's Dangerous Idea.* Allen Lane: The Penguin Press, London.

Dworkin R. 1981. "What is equality? Parts I and II." *Philosophy and Public Affairs*, 10:185-345.

Elster J. 1992. *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens.* Russell Sage Foundation, New York.

Grafen A. 1979. "The hawk-dove game played between relatives." *Animal Behavior*, 27:905-907.
– 1984. "Natural selection, kin selection and group selection." In J. Krebs and N. Davies, editors, *Behavioural Ecology (Second Edition).* Blackwell, Oxford.

Hamilton W. 1963. "The evolution of altruistic behavior." *Amarican Naturalist*, 97:354-356.
– 1964. "The genetic evolution of social behavior, parts I and II." *Journal of Theoretical Biology*, 7:1-52.

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

Hughes A. 1988. *Evolution and Human Kinship.* Oxford University Press, Oxford.

Hume D. 1978. *A Treatise of Human Nature. 2nd edition.* Clarendon press, Oxford. (Edited by L.A. Selbye-Bigge, revised by P. Nidditch, first published 1739.).

Humphrey N. 1976. "The social function of intellect." In P. Bateson and R. Hinde, editors, *Growing Points in Ethology.* Cambridge University Press, London.

Huxley T. 1989. *Evolution and Ethics: with New Essays on its Victorian and Sociobiological Context.* Princeton University Press, Princeton. (Edited by J. Paradis and G. Williams: first published 1893).

Kant I. 1964. *Groundwork of the Metaphysic of Morals.* Harper, New York. (Translated and analyzed by H. Paton. First published 1785.).

Krebs J. and N. Davies 1984. *Behavioural Ecology* (Second Edition). Blackwell, Oxford.

Lumsden C. and E. Wilson 1981. *Genes, Mind and Culture.* Harvard University Press, Cambridge, Mass.

Maynard Smith J. 1982. *Evolution and the Theory of Games.* Cambridge University Press, Cambridge.

Moore G.E. 1988. *Principia Ethica.* Prometheus Books, Buffalo, N.Y.. (First published 1902.).

Nunney L. 1985. "Group selection, altruism and structured-deme models." *American Naturalist*, 126:212-230.

Pinker S. 1994. *The Language Instinct: The New Science of Language and Mind.* Penguin Books, London.

Rawls J. 1972. *A Theory of Justice.* Oxford University Press, Oxford.

Ridley M. 1993. *The Red Queen: Sex and the Evolution of Human Nature.* Viking: Penguin Books, London.

Singer P. 1980. *The Expanding Circle: Ethics and Sociobiology.* Farrar, Strauss and Giroux, New York .

Tooby J. 1987. "The emergence of evolutionary phychology." In D. Pines, editor, *Emerging Syntheses in Science.* Santa Fe Institute, Santa Fe.

Trivers R. 1971. "The evolution of reciprocal altruism." *Quarterly Review of Biology*, 46:35-56.
– 1985. *Social Evolution.* Benjamin Cummings, Menlo Park, California.

Wade M. 1985. Soft selection, hard selection, kin selection and group selection. *American Naturalist*, 125:61-73.

Williams G. 1974. *Adaption and Natural Selection.* Princeton University Press, Princeton.

Wilson D. 1983. "The group selection controversy: History and current status." *Annual Review of*

Ecology and Systematics, 14:159-187.
Wilson D. and E. Sober 1988. "Reviving the superorganism." *Journal of Theoretical Biology*, 136:337-356.
Wilson E. 1975. *Sociobiology: The New Synthesis.* Harvard University Press, Cambridge, Mass.
Wright R. 1994. *The Moral Animal.* Random House, New York.
Wynne-Edwards V. 1962. *Animal Dispresion in Relation to Social Behavior.* Oliver and Boyd, Edinburgh.
Young P. 1994. *Equity.* Princeton University Press, Princeton.