Journal of Consciousness Studies
www.imprint-academic.com/jcs

Elliott Sober and David Sloan Wilson

Summary of: 'Unto Others

The Evolution and Psychology of Unselfish Behavior'

The hypothesis of group selection fell victim to a seemingly devastating critique in 1960s evolutionary biology. In Unto Others *(1998), we argue to the contrary, that group selection is a conceptually coherent and empirically well documented cause of evolution. We suggest, in addition, that it has been especially important in human evolution. In the second part of* Unto Others*, we consider the issue of psychological egoism and altruism — do human beings have ultimate motives concerning the well-being of others? We argue that previous psychological and philosophical work on this question has been inconclusive. We propose an evolutionary argument for the claim that human beings have altruistic ultimate motives.*

I: Introduction

Part One of *Unto Others* (Sober & Wilson, 1998) addresses the biological question of whether evolutionary altruism exists in nature and, if so, how it should be explained. Part Two concerns the psychological question of whether any of our ultimate motives involves an irreducible concern for the welfare of others. Both questions are descriptive, not normative. And neither, on the surface, even mentions the topic of morality. How, then, do these evolutionary and psychological matters bear on issues about morality? And what relevance do these descriptive questions have for normative ethical questions? These are problems we'll postpone discussing until we have outlined the main points we develop in *Unto Others*.

A behaviour is said to be altruistic in the evolutionary sense of that term if it involves a fitness cost to the donor and confers a fitness benefit on the recipient. A mindless organism can be an evolutionary altruist. It is important to recognize that the costs and benefits that evolutionary altruism involves come in the currency of reproductive success. If we give you a package of contraceptives as a gift, this won't be evolutionarily altruistic if the gift fails to enhance your reproductive success. And parents who take care of their children are not evolutionarily altruistic if they rear more children to adulthood than do parents who neglect their children. Evolutionary altruism is not the same as helping.

The concept of psychological altruism is, in a sense, the mirror image of the evolutionary concept. Evolutionary altruism describes the fitness effects of a behaviour,

Journal of Consciousness Studies, **7**, No. 1–2, 2000, pp. 185–206

not the thoughts or feelings, if any, that prompt individuals to produce those behaviours. In contrast, psychological altruism concerns the motives that cause a behavior, not its actual effects. If your treatment of others is prompted by your having an ultimate, noninstrumental concern for their welfare, this says nothing as to whether your actions will in fact be beneficial. Similarly, if you act only to benefit yourself, it is a further question what effect your actions will have on others. Psychological egoists who help because this makes them feel good may make the world a better place. And psychological altruists who are misguided, or whose efforts miscarry, can make the world worse.

Although the two concepts of altruism are distinct, they often are run together. People sometimes conclude that if genuine evolutionary altruism does not exist in nature, then it would be mere wishful thinking to hold that psychological altruism exists in human nature. The inference does not follow.

II: Evolutionary Altruism — Part One of *Unto Others*

1. The problem of evolutionary altruism and the critique of group selection in the 1960s

Evolutionary altruism poses a fundamental problem for the theory of natural selection. By definition, altruists have lower fitness than the selfish individuals with whom they interact. It therefore seems inevitable that natural selection should eliminate altruistic behaviour, just as it eliminates other traits that diminish an individual's fitness. Darwin saw this point, but he also thought that he saw genuinely altruistic characteristics in nature. The barbed stinger of a honey bee causes the bee to die when it stings an intruder to the nest. And numerous species of social insects include individual workers who are sterile. In both cases, the trait is good for the group though deleterious for the individuals who have it. In addition to these examples from nonhuman species, Darwin thought that human moralities exhibit striking examples of evolutionary altruism. In *The Descent of Man*, Darwin (1871) discusses the behaviour of courageous men who risk their lives to defend their tribes when a war occurs. Darwin hypothesized that these characteristics cannot be explained by the usual process of natural selection in which individuals compete with other individuals in the same group. This led him to advance the hypothesis of *group selection*. Barbed stingers, sterile castes, and human morality evolved because groups competed against other groups. Evolutionarily selfish traits evolve if selection occurs exclusively at the individual level. Group selection makes the evolution of altruism possible.

Although Darwin invoked the hypothesis of group selection only a few times, his successors were less abstemious. Group selection became an important hypothesis in the evolutionary biologist's toolkit during the heyday of the Modern Synthesis (c. 1930–1960). Biologists invoked individual selection to explain some traits, such as sharp teeth and immunity to disease; they invoked group selection to explain others, such as pecking order and the existence of genetic variation within species. Biologists simply used the concept that seemed appropriate. Discussion of putative group adaptations were not grounded in mathematical models of the group selection process, which hardly existed. Nor did naturalists usually feel the need to supply a mathematical model to support the claim that this or that phenotype evolved by individual selection.

All this changed in the 1960s when the hypothesis of group selection was vigorously criticized. It was attacked not just for making claims that are empirically false, but for being conceptually confused. The most influential of these critiques was George C. Williams' 1966 book, *Adaptation and Natural Selection*. Williams argued that traits don't evolve because they help groups; and even the idea that they evolve because they benefit individual organisms isn't quite right. Williams proposed that the right view is that traits evolve because they promote the replication of genes.

Williams' book, like much of the literature of that period, exhibits an ambivalent attitude towards the idea of group selection. Williams was consistently against the hypothesis; what he was ambivalent about was the grounds on which he thought the hypothesis should be rejected. Some of Williams' book deploys empirical arguments against group selection. For example, he argues that individual selection and group selection make different predictions about the sex ratio (the proportion of males and females) that should be found in a population; he claimed that the observations are squarely on the side of individual selection. But a substantial part of Williams' book advances somewhat *a priori* arguments against group selection. An example is his contention that the gene is the unit of selection because genes persist through many generations, whereas groups, organisms, and gene complexes are evanescent. Another example is his contention that group selection hypotheses are less parsimonious than hypotheses of individual selection, and so should be rejected on that basis.

The attack on group selection in the 1960s occurred at the same time that new mathematical models made it seem that the hypothesis of group selection was superfluous. W.D. Hamilton published an enormously influential paper in 1964, which begins with the claim that the classical notion of Darwinian fitness — an organism's prospects of reproductive success — can explain virtually none of the helping behaviour we see in nature. It can explain parental care, but when individuals help individuals who are not their offspring, a new concept of fitness is needed to explain why. This led Hamilton to introduce the mathematical concept of *inclusive fitness*. The point of this concept was to show how helping a relative and helping one's offspring can be brought under the same theoretical umbrella — both evolve because they enhance the donor's inclusive fitness. Many biologists concluded that helping behaviour directed at relatives is therefore an instance of selfishness, not altruism. Helping offspring and helping kin are both in one's genetic self-interest, because both allow copies of one's genes to make their way into the next generation. Behaviours that earlier seemed instances of altruism now seemed to be instances of genetic selfishness. The traits that Darwin invoked the hypothesis of group selection to explain apparently can be explained by 'kin selection' (the term that Maynard Smith, 1964, suggested for the process that Hamilton described), which was interpreted as an instance of individual selection. Group selection wasn't needed as a hypothesis; it was 'unparsimonious'.

Another mathematical development that pushed group selection further into the shadows was evolutionary game theory. Maynard Smith, one of the main architects of evolutionary game theory, wanted to provide a sane alternative to sloppy group selection thinking. Konrad Lorenz and others had suggested, for example, that animals restrain themselves in intraspecific combat because this is good for the species. Maynard Smith and Price (1973) developed their game of hawks versus doves to show how restraint in combat can result from purely individual selection. Each individual in the population competes with one other individual, chosen at random, to

determine which will obtain some fitness benefit. Each plays either the hawk strategy of all-out fighting or the dove strategy of engaging in restrained and brief aggression. When a hawk fights a hawk, one of them gets the prize, but each stands a good chance of serious injury or death. When a hawk fights a dove, the hawk wins the prize and the dove beats a hasty retreat, thus avoiding serious injury. And when two doves fight, the battle is over quickly; there is a winner and a loser, but neither gets hurt. In this model, which trait does better depends on which trait is common and which is rare. If hawks are very common, a dove will do better than the average hawk — the average hawk gets injured a lot, but the dove does not. On the other hand, if doves are very common, a hawk will do better than the average dove. The evolutionary result is a polymorphism. Neither trait is driven to extinction; both are represented in the population. What Lorenz tried to explain by invoking the good of the species, Maynard Smith and Price proposed to explain purely in terms of individual advantage. Just as was true in the case of Hamilton's work on inclusive fitness, the hypothesis of group selection appeared superfluous. You don't *need* the hypothesis to explain what you observe. Altruism is only an appearance. Dovishness isn't present because it helps the group; the trait is maintained in the population because individual doves gain an advantage from not fighting to the death.

Another apparent nail in the coffin of group selection was Maynard Smith's (1964) 'haystack model' of group selection. Maynard Smith considered the hypothetical situation in which field mice live in haystacks. The process begins by fertilized females each finding their own haystacks. Each gives birth to a set of offspring who then reproduce among themselves, brothers and sisters mating with each other. After that, the haystack holds together for another generation, with first cousins mating with first cousins. Each haystack contains a group of mice founded by a single female that sticks together for some number of generations. At a certain point, all the mice come out of their haystacks, mate at random, and then individual fertilized females go off to found their own groups in new haystacks. Maynard Smith analyzed this process mathematically and concluded that altruism can't evolve by group selection. Group selection is an inherently weak force, unable to overcome the countervailing and stronger force of individual selection, which promotes the evolution of selfishness.

The net effect of the critique of group selection in the 1960s was that the existence of adaptations that evolve because they benefit the group was dismissed from serious consideration in biology. The lesson was that the hypothesis of group selection doesn't have to be considered as an empirical possibility when the question is raised as to why this or that trait evolved. You know *in advance* that group selection is not the explanation. Only those who cling to the illusion that nature is cuddly and hospitable could take the hypothesis of group adaptation seriously.

2. Conceptual arguments against group selection

In *Unto Others*, we argue that this seemingly devastating critique of group selection completely missed the mark. The purely conceptual arguments against group selection show nothing. And the more empirical arguments also are flawed.

Let us grant that genes — not organisms or groups of organisms — are the *units of replication*. By this we mean that they are the devices that insure heredity. Offspring resemble parents because genes are passed from the latter to the former. However, this establishes nothing about why the adaptations found in nature have evolved.

Presumably, even if the gene is the unit of replication, it still can be true that some genes evolve because they code for traits that benefit individuals — this is why sharp teeth and immunity from disease evolve. But the same point holds for groups: even if the gene is the unit of replication, it remains to be decided whether some genes evolve because they code for traits that benefit groups. The fact that genes are *replicators* is entirely irrelevant to the units of *selection* problem.

The idea that group selection should be rejected because it is unparsimonious also fails to pass muster. Here's an example of how the argument is deployed, in Williams (1966), in Dawkins (1976), and in many other places. Why do crows exhibit sentinel behaviour? Group selection was sometimes invoked to explain this as an instance of altruism. A crow that sights an approaching predator and issues a warning cry places itself at risk by attracting the predator's attention; in addition, the sentinel confers a benefit on the other crows in the group by alerting them to danger. Interpreted in this way, a group selection explanation may seem plausible. However, an alternative possibility is that the sentinel behaviour is not really altruistic at all. Perhaps the sentinel cry is difficult for the predator to locate, and maybe the cry sends the other crows in the group into a frenzy of activity, thus permitting the sentinel to beat a safe retreat. If the behaviour is selfish, no group selection explanation is needed. At this point, one might think that two empirical hypotheses have been presented and that observations are needed to test which is better supported. However, the style of parsimony argument advanced in the anti-group selection literature concludes without further ado that the group selection explanation should be rejected, just because an individual selection explanation has been *imagined*. Data aren't needed, because parsimony answers our question. In *Unto Others*, we argue that this is a spurious application of the principle of parsimony. Parsimony is a guide to how observations should be interpreted; it is not a substitute for performing observational tests.

There is another fallacy that has played a central role in the group selection debate. The fallacy involves defining 'individual selection' so that any trait that evolves because of selection is automatically said to be due to individual selection; the hypothesis that traits might evolve by group selection thus becomes a definitional impossibility. In *Unto Others*, we call this *the averaging fallacy.* To explain how the fallacy works, let's begin with the standard representation of fitness payoffs to altruistic (A) and selfish (S) individuals when they interact in groups of size two. The argument would not be different if we considered larger groups. When two individuals interact, the payoff to the row player depends on whether he is A or S and on whether the person he interacts with is A or S (b is the benefit to the recipient and c is the cost to the altruistic A-type's behaviour):

		the other player is	
		A	S
fitness of a	A	$x + b - c$	$x - c$
player who is	S	$x + b$	x

What is the average fitness of A individuals? It will be an average — an altruist has a certain probability (p) of being paired with another altruist, and the complementary probability ($1-p$) of being paired with a selfish individual. Likewise, a selfish individual has a certain probability of being paired with an altruist (q) and the complementary probability ($1-q$) of being paired with another selfish individual. Thus, the fitnesses of the two traits are

$$w(A) = p(x+b-c) + (1-p)(x-c) = pb + x - c$$
$$w(S) = (q)(x+b) + (1-q)(x) = qb + x.$$

By definition, the trait with the higher average fitness will increase in frequency, if natural selection governs the evolutionary process. The criterion for which trait evolves is therefore:

(1) $w(A)>w(S)$ if and only if $p-q>c/b$.

The quantity $(p-q)$, we emphasize, is the difference between two probabilities:

$(p-q)$ = the probability that an altruist has of interacting with another altruist minus the probability that a selfish individual has of interacting with an altruist.

This difference represents the *correlation* of the two traits.

Two consequences of proposition (1) are worth noting:

(2) When like interacts with like, $w(A)>w(S)$ if and only if $b>c$.

(3) When individuals interact at random, $w(A)>w(S)$ if and only if $0>c/b$.

Proposition (2) identifies the case most favourable for the evolution of A — as long as the benefit to the recipient is greater than the cost incurred by the donor, A will evolve. Proposition (3), on the other hand, describes a situation in which A cannot evolve, as long as c and b are both greater than zero.

This analysis of the evolutionary consequences of the payoffs stipulated for traits A and S is not controversial. The fallacy arises when it is proposed that the selfish trait is the trait that has the higher average fitness, and that individual selection is the process that causes selfishness, so defined, to evolve. The effect of this proposal is that A is said to be selfish in situation (2) if $b>c$, while S is labelled selfish in situation (3), if $b,c>0$. Selfishness is equated with 'what evolves', and individual selection is, by definition, the selection process that makes selfishness evolve. This framework entails that altruism cannot evolve by natural selection and that group selection cannot exist. We reject this definitional framework because it fails to do justice to the biological problem that Darwin and his successors were addressing. The question of what types of adaptations are found in nature is *empirical*. If altruism and group adaptations do not exist, this must be demonstrated by observation. The real question cannot be settled by this semantic sleight of hand.

Our proposal is to define altruism and selfishness by the payoff matrix given above. What is true, by definition, is that altruists are less fit than selfish individuals *in the same group*. If b and c are both positive, then $x+b>x-c$. However, nothing follows from this as to whether altruists have lower fitness when one averages *across all groups*. This will be not be the case in the circumstance described in proposition (2), if $b>c$, but will be the case in the situation described in (3), if $b,c>0$.

Given the payoffs described, groups vary in fitness; the average fitness in AA groups is *(x+b-c)*, the average fitness in AS groups is *(x + [b-c]/2)*, and the average fitness in SS groups is *x*. Group selection favours altruism; groups do better the more altruists they contain. Individual selection, on the other hand, favours selfishness. There is no individual selection within homogeneous groups; the only individual (i.e., within-group) selection that occurs is in groups that are AS. Within such groups, selfishness outcompetes altruism. Here group and individual selection are opposing forces; which force is stronger determines whether altruism increases or declines in frequency in the ensemble of groups. Just as Darwin conjectured, it takes group selection for altruism to evolve.

Our proposal — that altruism and selfishness should be defined by the payoff matrix described above, and that group selection involves selection among groups, whereas individual selection involves selection within groups — is not something we invented, but reflects a long-standing set of practices in biology. Fitness averaged across groups is a criterion for which trait evolves. However, if one additionally wants to know whether group selection is part of the process, one must decompose this average by making within-group and between-group fitness comparisons.

This perspective on what altruism and group selection mean undermines the pervasive opinion that kin selection and game-theoretic interactions are alternatives to group selection. It also allows us to re-evaluate Hamilton's claim that classical Darwinian fitness cannot explain the evolution of helping behaviour (other than that of parental care) and that the concept of inclusive fitness is needed. The inclusive fitness of altruism reflects the cost to the donor and the benefit to the recipient, the latter weighed by the coefficient of relatedness (r) that donor bears to recipient:

$$I(A) = x - c + br.$$

The inclusive fitness of a selfish individual is

$$I(S) = x.$$

Notice that I(A) does not reflect the possibility that the altruist in question may receive a donation from another altruist, and the same is true of I(S) — it fails to reflect the possibility that a selfish individual may receive a donation from an altruist. The reason for these omissions is that we are assuming that altruism is *rare*. In any event, from these two inclusive fitnesses, we obtain 'Hamilton's rule' for the evolution of altruism:

$$(4) \quad r>c/b.$$

We hope the reader notices a resemblance between propositions (1) and (4). The coefficient of relatedness is a way of expressing the correlation of interactors. Contrary to Hamilton (1964), the concept of inclusive fitness is *not* needed to describe the circumstances in which altruism will evolve.

The coefficient of relatedness '*r*' is relevant to the evolution of altruism because related individuals tend to resemble each other. What is crucial for the evolution of altruism is that altruists tend to interact with altruists. This can occur because relatives tend to interact with each other, or because unrelated individuals who resemble each other tend to interact. The natural conclusion to draw is that kin selection is a kind of group selection, in which the groups are composed of relatives. When an

altruistic individual helps a related individual who is selfish, the donor still has a lower fitness than the recipient. The fact that they are related does not cancel this fundamental fact. *Within* a group of relatives, altruists are less fit than selfish individuals. It is only because of selection *among* groups that altruism can evolve. This, by the way, is the interpretation that Hamilton (1975) himself embraced about his own work, but his changed interpretation apparently has not been heard by many of his disciples.

Similar conclusions need to be drawn about game theory. Perhaps the most famous study in evolutionary game theory is the set of simulations carried out by Axelrod (1984). Axelrod had various game theorists suggest strategies that individuals might follow in repeated interactions. Individuals pair up at random and then behave altruistically or selfishly towards each other on each of several interactions. The payoffs that come from each interaction are the ones described before. However, the situation is more complex because there are many strategies that individuals might follow. Some strategies are *unconditional* — for example, an individual might act selfishly on every move (ALLS) or it might act altruistically on every move. In addition, there are many *conditional* strategies, according to which a player's action at one time depends on what has happened earlier in his interactions with the other player. Axelrod found that the strategy suggested by Anatol Rappaport of Tit-for-Tat (TFT) did better than many more selfish strategies. TFT is a strategy of *reciprocity.* A TFT player begins by acting altruistically and thereafter does whatever the other player did on the previous move. Two TFT players act altruistically towards each other on every move; if there are n moves in the game, each obtains a total score of $n(x-b+c)$. When TFT plays ALLS, the TFT player acts altruistically on the first move and then shifts to selfishness thereafter; if there are n interactions, TFT receives $(x-c) + (n-1)x = (nx-c)$ in its interaction with ALLS, who receives $(x+b) + (n-1)x$. Finally, if two ALLS players interact, each receives nx.

It is perfectly true, as a biographical matter, that Maynard Smith developed evolutionary game theory as an alternative to the hypothesis of group selection. However, the theory he described in fact involves group selection. If TFT competes with ALLS, there is group selection in which groups are formed at random and the groups are of size 2. Groups do better the more TFTers they contain. There is individual selection within mixed groups, in which TFT does worse than ALLS. TFT is able to evolve only because group selection favouring TFT overcomes the opposing force of individual (within-group) selection, which favours ALLS.

3. Empirical arguments against group selection

Williams (1966) proposed that sex ratio provides an empirical test of group selection. If sex ratio evolves by individual selection, then a roughly 1:1 ratio should be present. On the other hand, if sex ratio evolves by group selection, a female-biased sex ratio will evolve if this ratio helps the group to maximize its productivity. Williams then claims that the sex ratios found in nature are almost all close to even. He concludes that the case against group selection, with respect to this trait at least, is closed.

A year later, Hamilton (1967) reported that female-biased sex ratios are abundant. One might expect that the evolution community would have greeted Hamilton's report as providing powerful evidence in favour of group selection. This is exactly what did not occur. Although Hamilton described his own explanation of the

evolution of 'extraordinary sex ratios' as involving group selection, this is not how most other biologists interpreted it. Williams' sound reasoning that individual selection should produce an even sex ratio traces back to a model first informally proposed by R.A. Fisher (1930). Fisher assumed that parents produce a generation of offspring; these offspring then mate with each other at random, thus producing the grandoffspring of the original parents. If the offspring generation is predominately male, then a parent does best by producing all daughters; if the offspring generation is predominately female, the parent does best by producing all sons. Selection favours parents who produce the minority sex, and the population evolves towards an even sex ratio as a result. Hamilton introduced a change in assumptions. He considered the example of parasitic wasps who lay their eggs in hosts. One or more fertilized females lays eggs in a host; the offspring of these original foundresses mate with each other, after which they disperse to find new hosts and the cycle starts anew. The important point about Hamilton's model is that offspring in different hosts don't mate with each other.

Williams observed, correctly, that the way for a group to maximize its productivity is for it to have the smallest number of males that is necessary to insure that all females are fertilized. Group selection therefore favours a female-biased sex ratio, and this in fact is what Hamilton's model explains. The wasps in a host form a group, and groups with a female-biased sex ratio are more productive than groups in which the sex ratio is even. This is how Hamilton (1967, footnote 43) interprets his model, but most of his readers apparently did not. Rather, they construed Hamilton's model as describing individual selection; the reason is that Hamilton analyzed his model by calculating what the 'unbeatable strategy' is — that is, the strategy whose fitness is greater than the alternatives. This is the sex ratio strategy that will evolve. To automatically equate the unbeatable strategy with 'what evolves by individual selection' is to commit the averaging fallacy. Instead of considering what goes on within hosts as an instance of individual selection and differences among hosts as reflecting the action of group selection, the mistake is to meld these two processes together to yield a single summary statistic, which reflects the fitnesses of strategies averaged across groups. There is nothing wrong with obtaining this average if one merely wishes to say what trait will evolve. However, if the goal, additionally, is to say whether group selection is in part responsible for the evolutionary outcome, one can't use a framework in which what evolves is automatically equated with pure individual selection.

The other empirical argument we mentioned before, which was thought to tell against the hypothesis of group selection, is Maynard Smith's (1964) haystack model. It is a little odd to call this argument 'empirical', since it did not involve the gathering of data. Rather, the argument was 'theoretical', based on the analysis of a hypothetical model. In any event, let's consider how Maynard Smith managed to reach the conclusion that group selection is a weak force, unequal to the task of overcoming the opposing force of individual selection. The answer is that Maynard Smith simply *stipulated* that the within-haystack, individual selection part of his process was as powerful as it could possibly be. He *assumes without argument* that altruism is driven to extinction in all haystacks in which it is mixed with selfishness; the only way that altruism can survive in a haystack is by being in a haystack that is 100 per cent altruistic. We do not dispute that, as a matter of definition, altruism must *decline* in frequency in all mixed haystacks. But the idea that it must *decline to zero* in all such

haystacks is *not* a matter of definition. In effect, Maynard Smith explored a worse-case scenario for group selection. This tells us nothing as to whether altruism can evolve by group selection. Twenty years later, one of us (DSW) explored the question in a more general setting. The result is that altruism can evolve by group selection for a reasonable range of parameter values. The haystack model is not the stake through the heart of group selection that it was thought to be.

4. Multilevel selection theory is pluralistic

It is one thing to undermine fallacious arguments against group selection. It is something quite different to show that group selection has actually occurred and that it has been an important factor in the evolution of some traits. We attempt to do both in *Unto Others*. Sex ratio evolution is an especially well documented trait that has been influenced by group selection. But there are others — the evolution of reduced virulence in disease organisms, for example. Rather than discussing other examples, we want to make some general comments about the overall theory we are proposing.

First, our claim is not that *all* sex ratios in *all* populations are group adaptations. As Fisher argued, even sex ratios are plausibly regarded as individual adaptations. And as for the female-biased sex ratios found in nature, our claim is not that group selection was the *only* factor influencing their evolution. We do not claim that these groups have the smallest number of males consistent with all the females being fertilized. Rather, we claim that the biased sex ratios that evolve are *compromises* between the simultaneous and opposite influences of group and individual selection. Group selection rarely, if ever, occurs without individual selection occurring as well.

The more general point we want to emphasize is that hypotheses of group selection need to be evaluated on a trait-by-trait and a lineage-by-lineage basis. Group selection influenced sex ratio in some species, but not in others. And the fact that group selection did not influence sex ratio in human beings, for example, leaves open the question of whether group selection has been an important influence on other human traits. Unlike the monolithic theory of the selfish gene, which claims that *all* traits in *all* lineages evolved for the good of the genes, the theory we advocate, *multilevel selection theory,* is pluralistic. Different traits evolved because of different combinations of causes.

5. Group selection and human evolution

In *Unto Others*, we develop the conjecture that group selection was a strong force in human evolution. Group selection includes, but is not confined to, direct intergroup competition such as warfare. But, just as individual plants can compete with each other in virtue of the desert conditions in which they live (some being more drought-resistant than others), so groups can compete with each other without directly interacting (e.g., by some groups fostering co-operation more than others). In addition, cultural variation in addition to genetic variation can provide the mechanisms for phenotypic variation and heritability at the group level (see also Boyd and Richerson, 1985).

As noted earlier, the evolution of altruism depends on altruists interacting preferentially with each other. Kin selection is a powerful idea because interaction among kin is a pervasive pattern across many plant and animal groups. However, in many organisms, including especially human beings, individuals *choose* the individuals

with whom they interact. If altruists seek out other altruists, this promotes the evolution of altruism. Although kin selection is a kind of group selection, there can be group selection that isn't kin selection; this, we suspect, is especially important in the case of human evolution. However, it isn't *uniquely* human — for example, even so-called lower vertebrates such as guppies can choose the social partners with which they interact.

An additional factor that helps altruism to evolve, which may be uniquely human, is the existence of cultural norms that impose social controls. Consider a very costly act, such as donating ten per cent of your food to the community. Since this act is very costly, a very strong degree of correlation among interactors will be needed to get it to evolve. However, suppose you live in a society in which individuals who make the donation are rewarded, and those who do not are punished. The act of donation has been transformed. It is no longer altruistic to make the donation, but selfish. Individuals in your group who donate do better than individuals who do not. However, it would be wrong to conclude from this that the existence of social controls make the hypothesis of group selection unnecessary. For where did the existence and enforcement of the social sanctions come from? Why do some individuals enforce the penalty for nondonation? This costs them something. A free-rider could enjoy the benefits without paying the costs of having a norm of donation enforced. Enforcing the requirement of donation is altruistic, even if donation is no longer altruistic. But notice that the cost of being an enforcer may be slight. It may not cost you anything like ten per cent of your food supply to help enforce the norm of donation. This means that the degree of correlation among interactors needed to get *this* altruistic behaviour to evolve is much less.

We believe that this argument may explain how altruistic behaviours were able to evolve in the genetically heterogeneous groups in which our ancestors lived. Human societies, both ancient and modern, are nowhere near as genetically uniform as bee hives and ant colonies. How, then, did co-operative behaviour manage to evolve in them? Human beings, we believe, did something that no other species was able to do. Social norms convert highly altruistic traits into traits that are selfish. And enforcing a social norm can involve a smaller cost than the required behaviour would have imposed if there were no norms. Social norms allow social organization to evolve by reducing its costs. Here again, it is important to recognize that culture allows a form of selection to occur whose elements may be found in the absence of culture. Bees 'police' the behaviour of other bees. What is uniquely human is the harnessing of socially shared values.

In addition to these rather 'theoretical' considerations, *Unto Others* also presents some observations that support the hypothesis that human beings are a group selected species. We randomly sampled twenty-five societies from the Human Relations Area File, an anthropological database, consulting what the files say about social norms. The actual contents of these norms vary enormously across our sample — for example, some societies encourage innovation in dress, while others demand uniformity. In spite of this diversity, cultural norms almost always require individuals to avoid conflict with each other and to behave benevolently towards fellow group members. Such constraints are rarely present with respect to outsiders, however. It also was striking how closely individuals can monitor the behaviour of group members in most traditional societies. Equally impressive is the emphasis on egalitarianism (among

males — not, apparently, between males and females) found in many traditional societies; the norm was not that there should be complete equality, but that inequalities are permitted only when they enhance group functioning.

In addition to this survey data, we also describe a 'smoking gun' of cultural group selection — the conflict between the Nuer and Dinka tribes in East Africa. This conflict has been studied extensively by anthropologists for most of this century. The Nuer have gradually eroded the territory and resources of the Dinka, owing to the Nuer's superior group organization. The transformation was largely underwritten by people in Dinka villages defecting to the Nuers and being absorbed into their culture. We conjecture that this example has countless counterparts in the human past, and that the process of cultural group selection that it exemplifies has been an important influence on cultural change.

We think that Part I of *Unto Others* provides a solid foundation for the theory of group selection and that we have presented several well-documented cases of group selection in nonhuman species. Our discussion of human group selection is more tentative, but nonetheless we are prepared to claim that human beings have been strongly influenced by group selection processes.

III: Psychological Altruism — Part Two of *Unto Others*

Psychological egoism is a theory that claims that all of our ultimate desires are self-directed. Whenever we want others to do well (or badly), we have these other-directed desires only instrumentally; we care about what happens to others only because we think that the welfare of others has ramifications for ourselves. Egoism has exerted a powerful influence in the social sciences and has made large inroads in the thinking of ordinary people. In Part Two of *Unto Others*, we review the philosophical and psychological arguments that have been developed about egoism, both *pro* and *con*. We contend that these arguments are inconclusive. A new approach is needed; in Chapter 10, we present an evolutionary argument for thinking that some of our ultimate motives are altruistic.

It is easy to invent egoistic explanations for even the most harrowing acts of self-sacrifice. The soldier in a foxhole who throws himself on a grenade to save the lives of his comrades is a fixture in the literature on egoism. How could this act be a product of self-interest, if the soldier knows that it will end his life? The egoist may answer that the soldier realizes in an instant that he would rather die than suffer the guilt feelings that would haunt him if he saved himself and allowed his friends to perish. The soldier prefers to die and have no sensations at all rather than live and suffer the torments of the damned. This reply may sound *forced*, but this does not show that it must be *false*. And the fact that an egoistic explanation can be *invented* is no sure sign that egoism is *true*.

1. Clarifying egoism

When egoism claims that all our ultimate desires are self-directed, what do 'ultimate' and 'self-directed' mean?

There are some things that we want for their own sakes; other things we want only because we think they will get us something else. The crucial relation that we need to define is this:

S wants *m* solely as a means to acquiring *e* if and only if *S* wants *m*, *S* wants *e,* and *S* wants *m* only because she believes that obtaining *m* will help her obtain *e*.

An ultimate desire is a desire that someone has for reasons that go beyond its ability to contribute instrumentally to the attainment of something else. Consider pain. The most obvious reason that people want to avoid pain is simply that they dislike experiencing it. Avoiding pain is one of our ultimate goals. However, many people realize that being in pain reduces their ability to concentrate, so they may sometimes take an aspirin in part because they want to remove a source of distraction. This shows that the things we want as ends in themselves we also may want for instrumental reasons.

When psychological egoism seeks to explain why one person helped another, it isn't enough to show that *one* of the reasons for helping was self-benefit; this is quite consistent with there being another, purely altruistic, reason that the individual had for helping. Symmetrically, to refute egoism, one need not cite examples of helping in which *only* other-directed motives play a role. If people sometimes help for both egoistic and altruistic ultimate reasons, then psychological egoism is false.

Egoism and altruism both require the distinction between self-directed and other-directed desires, which should be understood in terms of a desire's propositional content. If Adam wants the apple, this is elliptical for saying that Adam wants it to be the case that *he has the apple*. This desire is purely self-directed, since its propositional content mentions Adam, but no other agent. In contrast, when Eve wants *Adam to have the apple*, this desire is purely other-directed; its propositional content mentions another person, Adam, but not Eve herself. Egoism claims that all of our ultimate desires are self-directed; altruism, that some are other-directed.

A special version of egoism is psychological hedonism. The hedonist says that the only ultimate desires that people have are attaining pleasure and avoiding pain. Hedonism is sometimes criticized for holding that pleasure is a single type of sensation — that the pleasure we get from the taste of a peach and the pleasure we get from seeing those we love prosper somehow boil down to the same thing (Lafollette, 1988). However, this criticism does not apply to hedonism as we have described it. The salient fact about hedonism is its claim that people are *motivational solipsists*; the only things they care about ultimately are states of their own consciousness. Although hedonists must be egoists, the reverse isn't true. For example, if people desire their own survival as an end in itself, they may be egoists, but they are not hedonists.

Some desires are neither purely self-directed nor purely other-directed. If Phyllis wants to be famous, this means that she wants others to know who she is. This desire's propositional content involves a relation between self and others. If Phyllis seeks fame solely because she thinks this will be pleasurable or profitable, then she may be an egoist. But what if she wants to be famous as an end in itself? There is no reason to cram this possibility into either egoism or altruism. So let us recognize *relationism* as a possibility distinct from both. Construed in this way, egoism avoids the difficulty of having to explain why the theory is compatible with the existence of some relational ultimate desires, but not with others (Kavka, 1986).

With egoism characterized as suggested, it obviously is not entailed by the truism that people act on the basis of their own desires, nor by the truism that they seek to have their desires satisfied. The fact that Joe acts on the basis of Joe's desires, not on the basis of Jim's, tells us *whose* desires are doing the work; it says nothing about whether the ultimate desires in Joe's head are *purely self-directed*. And the fact that

Joe wants his desires to be satisfied means merely that he wants their propositional contents to come true (Stampe, 1994). If Joe wants it to rain tomorrow, then his desire is satisfied if it rains, whether or not he notices the weather. To want one's desires satisfied is not the same as wanting the feeling of satisfaction that sometimes accompanies a satisfied desire.

Egoism is sometimes criticized for attributing too much calculation to spontaneous acts of helping. People who help in emergency situations often report doing so 'without thinking' (Clark and Word, 1974). However, it is hard to take such reports literally when the acts involve a precise series of complicated actions that are well-suited to an apparent end. A lifeguard who rescues a struggling swimmer is properly viewed as having a goal and as selecting actions that advance that goal. The fact that she engaged in no ponderous and self-conscious calculation does not show that no means/end reasoning occurred. In any case, actions that really do occur without the mediation of beliefs and desires fall outside the scope of both egoism and altruism.

A related criticism is that egoism assumes that people are more rational than they really are. However, recall that egoism is simply a claim about the ultimate desires that people have. As such, it says nothing about how people decide what to do on the basis of their beliefs and desires. The assumption of rationality is no more a part of psychological egoism than it is part of *motivational pluralism* — the view that people have both egoistic and altruistic ultimate desires.

2. Psychological arguments

It may strike some readers that deciding between egoism and motivational pluralism is easy. Individuals can merely gaze within their own minds and determine by introspection what their ultimate motives are. The problem with this easy solution is that there is no independent reason to think that the testimony of introspection is to be trusted in this instance. Introspection is misleading or incomplete in what it tells us about other facets of the mind; there is no reason to think that the mind is an open book with respect to the issue of ultimate motives.

In *Unto Others*, we devote most of Chapter 8 to the literature in social psychology that seeks to test egoism and motivational pluralism experimentally. The most systematic attempt in this regard is the work of Batson and co-workers, summarized in Batson (1991). Batson tests a hypothesis he calls the *empathy-altruism hypothesis* against a variety of egoistic explanations. The empathy-altruism hypothesis asserts that empathy causes people to have altruistic ultimate desires. We argue that Batson's experiments succeed in refuting some simple forms of egoism, but that the perennial problem of refuting egoism remains — when one version of egoism is refuted by a set of observations, another can be invented that fits the data. We also argue that even if Batson's experiments show that empathy causes helping, they don't settle whether empathy brings about this result by triggering an altruistic ultimate motive. We don't conclude from this that experimental social psychology will never be able to answer the question of whether psychological egoism is true. Our negative conclusion is more modest — empirical attempts to decide between egoism and motivational pluralism have not yet succeeded.

3. A bevy of philosophical arguments

Egoism has come under fire in philosophy from a number of angles. In Chapter 9 of *Unto Others*, we review these arguments and conclude that none of them succeeds. Here, briefly, is a sampling of the arguments we consider, and our replies:

— Egoism has been said to be *untestable*, and thus not a genuine scientific theory at all. We reply that if egoism is untestable, so is motivational pluralism. If it is true that when one egoistic explanation is discredited, another can be invented in its stead, then the same can be said of pluralism. The reason that egoism and pluralism have this sort of flexibility is that both make claims about the *kinds* of explanations that human behaviour has; they do not provide a detailed explanation of any particular behaviour. Egoism and pluralism are *ism*s, which are notorious for the fact that they are not crisply falsifiable by a single set of observations.

— Joseph Butler (1692–1752) is widely regarded as having refuted psychological hedonism (Broad, 1965; Feinberg, 1984; Nagel, 1970). His argument can be outlined as follows:

1. People sometimes experience pleasure.
2. When people experience pleasure, this is because they had a desire for some external thing, and that desire was satisfied.

∴ Hedonism is false.

We think the second premise is false. It is overstated; although some pleasures are the result of a desire's being satisfied, others are not (Broad, 1965, p. 66). One can enjoy the smell of violets without having formed the desire to smell a flower, or something sweet. Since desires are propositional attitudes, forming a desire is a cognitive achievement. Pleasure and pain, on the other hand, are sometimes cognitively mediated, but sometimes they are not. This defect in the argument can be repaired; Butler does not need to say that desire satisfaction is the one and only road to pleasure. The main defect in the argument occurs in the transition from premises to conclusion. Consider the causal chain from a *desire* (the desire for food, say), to an *action* (eating), to a *result* — pleasure. Because the pleasure traces back to an antecedently existing desire, it will be false that the resulting pleasure caused the desire (on the assumption that cause must precede effect). However, this does not settle how two *desires* — the *desire for food* and the *desire for pleasure* — are related. Hedonism says that people desire food *because* they want pleasure (and think that food will bring them pleasure). Butler's argument concludes that this causal claim is false, but for no good reason. The crucial mistake in the argument comes from confusing two quite different items — the *pleasure* that results from a desire's being satisfied and the *desire for pleasure*. Even if the occurrence of pleasure presupposed that the agent desired something besides pleasure, nothing follows about the relationship between the *desire for pleasure* and the desire for something else (Sober, 1992; Stewart, 1992). Hedonism does not deny that people desire external things; rather, the theory tries to explain why that is so.

— We also consider the argument against egoism that Nozick (1974) presents by his example of an 'experience machine', the claim that hedonism is a paradoxical and irrational motivational theory, and the claim that egoism has the burden of proof. We conclude that none of these attacks on egoism is decisive.

There is one philosophical argument that attempts to support egoism, not refute it. This is the claim that egoism is preferable to pluralism because the former theory is more parsimonious. Egoism posits one type of ultimate desire whereas pluralism says there are two. We have two criticisms. First, this parsimony argument measures a theory's parsimony by counting the kinds of ultimate desires it postulates. The opposite conclusion would be obtained if one counted *causal beliefs*. The pluralist says that people want others to do well and that they also want to do well themselves. The egoist says that a person wants others to do well only because he or she *believes* that this will promote self-interest. Pluralism does not include this belief attribution. Our second objection is that parsimony is a reasonable tie-breaker when all other considerations are equal; it remains to be seen whether egoism and pluralism are equally plausible on all other grounds. In Chapter 10, we propose an argument to the effect that pluralism has greater evolutionary plausibility.

4. An evolutionary approach

Psychological motives are *proximate mechanisms* in the sense of that term used in evolutionary biology. When a sunflower turns towards the sun, there must be some mechanism inside the sunflower that causes it to do so. Hence, if phototropism evolved, a proximate mechanism that causes that behaviour also must have evolved. Similarly, if certain forms of helping behaviour in human beings are evolutionary adaptations, then the motives that cause those behaviours in individual human beings also must have evolved. Perhaps a general perspective on the evolution of proximate mechanisms can throw light on whether egoism or motivational pluralism was more likely to have evolved.

Pursuing this evolutionary approach does not presuppose that every detail of human behaviour, or every act of helping, can be explained completely by the hypothesis of evolution by natural selection. In Chapter 10, we consider a single fact about human behaviour, and our claim is that selection is relevant to explaining it. The phenomenon of interest is that human parents take care of their children; the average amount of parental care provided by human beings is strikingly greater than that provided by parents in many other species. We will assume that natural selection is at least part of the explanation of why parental care evolved in our lineage. This is not to deny that human parents vary; some take better care of their children than others, and some even abuse and kill their offspring. Another striking fact about individual variation is that mothers, on average, expend more time and effort on parental care than fathers. Perhaps there are evolutionary explanations for these individual differences as well; the question we want to address here, however, makes no assumption as to whether this is true.

In Chapter 10, we describe some general principles that govern how one might predict the proximate mechanism that will evolve to cause a particular behaviour. We develop these ideas by considering the example of a marine bacterium whose problem is to avoid environments in which there is oxygen. The organism has evolved a particular behaviour — it tends to swim away from greater oxygen concentrations and towards areas in which there is less. What proximate mechanism might have evolved that allows the organism to do this?

First, let's survey the range of possible design solutions that we need to consider. The most obvious solution is for the organism to have an oxygen detector. We call this

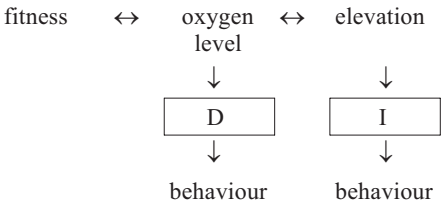
the *direct solution* to the design problem; the organism needs to avoid oxygen and it solves that problem by detecting the very property that matters.

It isn't hard to imagine other solutions to the design problem that are less direct. Suppose that areas near the pond's surface contain more oxygen and areas deeper in the pond contain less. If so, the organism could use an up/down detector to make the requisite discrimination. This design solution is *indirect*; the organism needs to distinguish high oxygen from low and accomplishes this by detecting another property that happens to be correlated with the target. In general, there may be many indirect design solutions that the organism could exploit; there are as many indirect solutions as there are correlations between oxygen level and other properties found in the environment. Finally, we may add to our list the idea that there can be *pluralistic* solutions to a design problem. In addition to the monistic solution of having an oxygen detector and the monistic solution of having an up/down detector, an organism might deploy both.

Given this multitude of possibilities, how might one predict which of them will evolve? Three principles are relevant — *availability*, *reliability*, and *efficiency*.

Natural selection acts only on the range of variation that exists ancestrally. An oxygen detector might be a good thing for the organism to have, but if that device was never present as an ancestral variant, natural selection cannot cause it to evolve. So the first sort of information we'd like to have concerns which proximate mechanisms were *available* ancestrally.

Let's suppose for the sake of argument that both an oxygen detector and an up/down detector are available ancestrally. Which of them is more likely to evolve? Here we need to address the issue of *reliability*. Which device does the more reliable job of indicating where oxygen is? Without further information, not much can be said. An oxygen detector may have any degree of reliability, and the same is true of an up/down detector. There is no *a priori* reason why the direct strategy should be more or less reliable than the indirect strategy. However, there is a special circumstance in which they will differ. It is illustrated by the following diagram:



The double arrows indicate correlation; avoiding oxygen is correlated with fitness, and elevation is correlated with oxygen level. In the diagram, there is no arrow from elevation to fitness except the one that passes through oxygen level. This means that elevation is correlated with fitness *only because* elevation is correlated with oxygen, and oxygen is correlated with fitness. There is no *a priori* reason why this should be true. For example, if there were more predators at the bottom of ponds than at the top, then elevation would have two sorts of relevance for fitness. However, if oxygen level 'screens off' fitness from elevation in the way indicated, we can state the following principle about the reliability of the direct device D and the indirect device I:

(D/I) If oxygen level and elevation are less than perfectly correlated, and if D detects oxygen level at least as well as I detects elevation, then D will be more reliable than I.

This is the Direct/Indirect Asymmetry Principle. Direct solutions to a design problem aren't always more reliable, but they are more reliable in this circumstance.

A second principle about reliability also can be extracted from this diagram. Just as scientists do a better job of discriminating between hypotheses if they have more evidence rather than less, so it will be true that the marine bacterium we are considering will make more reliable discriminations about where to swim if it has two sources of information rather than just one:

(TBO) If oxygen level and elevation are less than perfectly correlated, and if D and I are each reliable, though fallible, detectors of oxygen concentration, then D and I working together will be more reliable than either of them working alone.

This is the Two-is-Better-than-One Principle. It requires an assumption — that the two devices do not interfere with each other when both are present in an organism.

The D/I Asymmetry and the TBO Principle pertain to the issue of reliability. Let us now turn to the third consideration that is relevant to predicting which proximate mechanism will evolve, namely *efficiency*. Even if an oxygen detector and an elevation detector are both available, and even if the oxygen detector is more reliable, it doesn't follow that natural selection will favour the oxygen detector. It may be that an oxygen detector requires more energy to build and maintain than an elevation detector. Organisms run on energy no less than automobiles do. Efficiency is relevant to a trait's overall fitness just as much as its reliability is.

With these three considerations in hand, let's return to the problem of predicting which motivational mechanism for providing parental care is likely to have evolved in the lineage leading to human beings. The three motivational mechanisms we need to consider correspond to three different rules for selecting a behaviour in the light of what one believes:

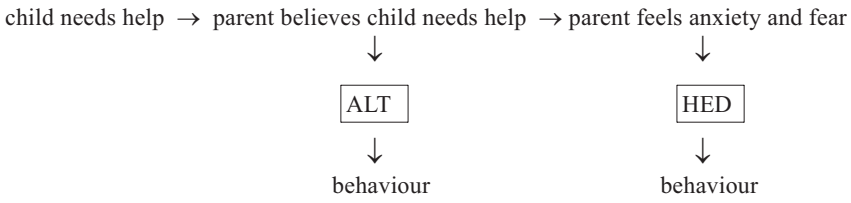
(HED) Provide parental care if, and only if, doing so will maximize pleasure and minimize pain.

(ALT) Provide parental care if, and only if, doing so will advance the welfare of one's children.

(PLUR) Provide parental care if, and only if, doing so will either maximize pleasure and minimize pain, or will advance the welfare of one's children.

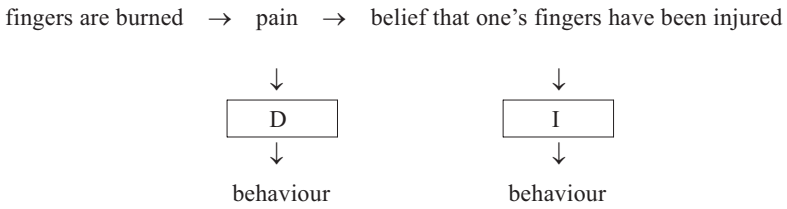
(ALT) is a relatively direct, and (HED) is a relatively indirect, solution to the design problem of getting an organism to take care of its offspring. Just as our marine bacterium can avoid oxygen by detecting elevation, so it is possible in principle for a hedonistic organism to provide parental care; what is required is that the organism be so constituted that providing parental care is the thing that usually maximizes its pleasure and minimizes its pain (or that the organism believes that this is so).

Let's consider how reliable these three mechanisms will be in a certain situation. Suppose that a parent learns that its child is in danger. Imagine that your neighbour tells you that your child has just fallen through the ice on a frozen lake. Here is how (HED) and (ALT) will do their work:



The altruistic parent will be moved to action just by virtue of believing that its child needs help. The hedonistic parent will not; rather, what moves the hedonistic parent to action are the feelings of anxiety and fear that are caused by the news. It should be clear from this diagram that the (D/I) Asymmetry Principle applies; (ALT) will be more reliable than (HED). And by the (TBO) Principle, (PLUR) will do better than both. In this example, hedonism comes in last in the three-way competition, at least as far as reliability is concerned.

The important thing about this example is that the feelings that the parent has are *belief mediated*. The only reason the parent *feels* anxiety and fear is that the parent *believes* that its child is in trouble. This is true of many of the situations that egoism and hedonism are called upon to explain, but it is not true of all. For example, consider the following situation in which pain is a direct effect, and belief a relatively indirect effect, of bodily injury:



In this case, hedonism is a direct solution to the design problem; it would be a poor engineering solution to have the organism be unresponsive to pain and to have it withdraw its fingers from the flame only after it forms a belief about bodily injury. In this situation, *belief is pain-mediated* and the (D/I) Asymmetry Principle explains why a hedonistic focus on pain makes sense. However, the same principle indicates what is misguided about hedonism as a design solution when *pain is belief-mediated*, which is what occurs so often in the context of parental care.

If hedonism is less reliable than both pure altruism and motivational pluralism, how do these three mechanisms compare when we consider their availability and efficiency? With respect to availability, we make the following claim: *if hedonism was available ancestrally, so was altruism*. The reason is that the two motivational mechanisms differ in only a modest way. Both require a belief/desire psychology. And both the hedonistic and the altruistic parent want their children to do well; the only difference is that the hedonist has this propositional content as an instrumental desire while the altruist has it as an ultimate desire. If altruism and pluralism did not evolve, this was not because they were unavailable as variants for selection to act upon.

What about efficiency? Does it cost more calories to build and maintain an altruistic or a pluralistic organism than it does to build and maintain a hedonist? We don't see why. What requires energy is building the hardware that implements a

belief/desire psychology. However, we doubt that it makes an energetic difference whether the organism has one ultimate desire rather than two. People with more beliefs apparently don't need to eat more than people with fewer. The same point seems to apply to the issue of how many, or which, ultimate desires one has.

In summary, pure altruism and pluralism are both more reliable than hedonism as devices for delivering parental care. And, with respect to the issues of availability and efficiency, we find no difference among these three motivational mechanisms. This suggests that natural selection is more likely to have made us motivational pluralists than to have made us hedonists.

From an evolutionary point of view, hedonism is a bizarre motivational mechanism. What matters in the process of natural selection is an organism's ability to survive and be reproductively successful. Reproductive success involves not just the production of offspring, but the survival of those offspring to reproductive age. So what matters is the survival of one's own body and the bodies of one's children. Hedonism, on the other hand, says that organisms care ultimately about the states of their own consciousness, and about that alone. Why would natural selection have led organisms to care about something that is peripheral to fitness, rather than have them set their eyes on the prize? If organisms were unable to conceptualize propositions about their own bodies and the bodies of their offspring, that might be a reason. After all, it might make sense for an organism to exploit the indirect strategy of deciding where to swim on the basis of elevation rather than on the basis of oxygen concentration, if the organism cannot detect oxygen. But if an organism is smart enough to form representations about itself and its offspring, this justification of the indirect strategy will not be plausible. The fact that we evolved from ancestors who were cognitively less sophisticated makes it unsurprising that avoiding pain and attaining pleasure are two of our ultimate goals. But the fact that human beings are able to form representations with so many different propositional contents suggests that evolution supplemented this list of what we care about as ends in themselves.

IV: Evolutionary Altruism, Psychological Altruism, and Ethics

The study of ethics has a *normative* and a *descriptive* component. Normative ethics seeks to say what is good and what is right; it seeks to identify what we are obliged to do and what we are permitted to do. Descriptive ethics, on the other hand, is neutral on these normative questions; it attempts to *describe* and *explain* morality as a cultural phenomenon, not *justify* it. How does morality vary within and across cultures, and through time? Are there moral ideas that constitute cultural universals? And how is one to explain this pattern of variation?

Although we think our work on evolutionary and psychological altruism bears on these questions, we also think that it is important not to blur the problems. Psychological altruism is not the same as morality. And an explanation of why human beings hold a moral principle is not, in itself, a justification (or a refutation) of that principle.

We say that psychological altruism is not the same as morality because individuals can have concerns about the welfare of specific others without their formulating those concerns in terms of ethical principles. A mother chimp may want her offspring to have some food, but this does not mean that she thinks that all chimps should be well-fed, or that all mothers should take care of their offspring. Egoistic and altruistic

desires are both desires about specific individuals. Having self-directed preferences is not sufficient for having a morality; the same goes for other-directed preferences.

Why, then, did morality evolve? People can have specific likes and dislikes without this producing a socially shared moral code. And if everyone dislikes certain things, what is the point of there being a moral code that says that those things should be shunned? If everyone hates sticking pins in their toes, what is the point of an ethic that tells people that it is wrong to stick pins in their toes? And if parents invariably love their children, what would be the point of having a moral principle that tells parents that they ought to love their children? Behaviours that people do spontaneously by virtue of their own desires don't need to have a moral code laid on top of them. The obvious suggestion is that the social function of morality is to get people to do things that they would not otherwise be disposed to do, or to strengthen dispositions that people already have in weaker forms. Morality is not a mere redundant overlay on the psychologically altruistic motives we may have.

Functionalism went out of style in anthropology and other social sciences in part because it was hard to see what feedback mechanism might make institutions persist or disappear. Even if religion promotes group solidarity, how would that explain the persistence of religion? The idea of selection makes this question tractable. We hope that *Unto Others* will allow social scientists to explore the hypothesis that morality is a group adaptation. We do not deny that moral principles have functioned as ideological weapons, allowing some individuals to prosper at the expense of others in the same group. However, the hypothesis that moralities sometimes persist and spread because they benefit the group is not mere wishful thinking. Darwin's idea that features of morality can be explained by group selection needs to be explored.

What, if anything, do the evolutionary and psychological issues we discuss in *Unto Others* contribute to normative theory? Every normative theory relies on a conception of human nature. Sometimes this is expressed by invoking the *ought implies can principle*. If people ought to do something, then it must be possible for them to do it. Human nature circumscribes what is possible. We do not regard human nature as unchangeable. In part, this is because evolution isn't over. Genetic and cultural evolution will continue to modify the capacities that people have. But if we want to understand the capacities that people *now* have, surely an understanding of our evolutionary past is crucial. One lesson that may flow from the evolutionary and psychological study of altruism is that prisoners' dilemmas are in fact rarer than many researchers suppose. Decision theory says that it is irrational to co-operate (to act altruistically) in one-shot prisoners' dilemmas. However, perhaps some situations that appear to third parties to be prisoners' dilemmas really are not. Payoffs are usually measured in dollars, or in other tangible commodities. But if people sometimes care about each other, and not just about money, they are not irrational when they choose to co-operate in such interactions. Narrow forms of egoism make such behaviours appear irrational. Perhaps the conclusion to draw is not that people *are* irrational, but that the assumption of egoism needs to be rethought.

References

Axelrod, Robert (1984), *The Evolution of Co-operation* (New York: Basic Books).

Batson, C. Daniel (1991), *The Altruism Question: Toward A Social-Psychological Answer* (Hillsdale, NJ: Lawrence Erlbaum Associates).

Boyd, Robert and Richerson, Peter (1985), *Culture and the Evolutionary Process* (Chicago: University of Chicago Press).
Broad, C.D (1965), *Five Types of Ethical Theory* (Totowa, NJ: Littlefield, Adams).
Butler, Joseph (1726), *Fifteen Sermons preached at the Rolls Chapel*, reprinted in part in *British Moralists*, Volume 1, ed. L.A. Selby-Bigge (New York: Dover Books, 1965; originally published Oxford: The Clarendon Press, 1897).
Clark, R.D. and Word, L.E. (1974), 'Where is the apathetic bystander? Situational characteristics of the emergency', *Journal of Personality and Social Psychology*, **29**, pp. 279–87.
Darwin, Charles (1871), *The Descent of Man and Evolution in Relation to Sex* (London: Murray).
Dawkins, Richard (1976), *The Selfish Gene* (New York: Oxford University Press).
Feinberg, J. (1984), 'Psychological egoism', in *Reason at Work*, ed. S. Cahn, P. Kitcher and G. Sher (San Diego, Calif.: Harcourt Brace and Jovanovich), pp. 25–35.
Fisher, Ronald (1930), *The Genetical Theory of Natural Selection* (New York: Dover, 1958).
Hamilton, W.D. (1964), 'The Genetical evolution of social behaviour I and II', *Journal of Theoretical Biology*, **7**, pp. 1–16, pp. 17–52.
Hamilton, W.D. (1967), 'Extraordinary sex ratios', *Science*, **156**, pp. 477–88.
Hamilton, W.D. (1975), 'Innate social aptitudes of man — an approach from evolutionary genetics', in *Biosocial Anthropology*, ed. R. Fox (New York: John Wiley) pp. 133–15 .
Kavka, Gregory (1986), *Hobbesian Moral and Political Theory* (Princeton, NJ: Princeton University Press).
Lafollette, Hugh (1988), 'The truth in psychological egoism', in *Reason and Responsibility*, 7th edition, ed. J. Feinberg (Belmont, Calif.: Wadsworth), pp. 500–7.
Maynard Smith, John (1964), 'Group selection and kin selection', *Nature*, **201**, pp. 1145–6.
Maynard Smith, John and Price, George (1973), 'The logic of animal conflict', *Nature*, **246**, pp.15–18.
Nagel, Thomas (1970), *The Possibility of Altruism* (Oxford: Oxford University Press).
Nozick, Robert (1974), *Anarchy, State, and Utopia* (New York: Basic Books).
Sober, Elliott (1992), 'Hedonism and Butler's stone', *Ethics*, **103**, pp. 97–103.
Sober, Elliott and Wilson, David Sloan (1998), *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press).
Stampe, Dennis (1994), 'Desire', in *A Companion to the Philosophy of Mind*, ed. S. Guttenplan (Cambridge, Mass.: Basil Blackwell), pp. 244–50.
Stewart, R.M. (1992), 'Butler's argument against psychological hedonism', *Canadian Journal of Philosophy*, **22**, pp. 211–21.
Williams, George C. (1966), *Adaptation and Natural Selection* (Princeton, NJ: Princeton University Press).