2

The evolution of morality

> Where it is in his own interest, every organism may reasonably be expected to aid his fellows. Where he has no alternative, he submits to the yoke of communal servitude. Yet given a full chance to act in his own interest, nothing but expediency will restrain him from brutalizing, from maiming, from murdering his brother, his mate, his parent, or his child. Scratch an "altruist" and watch a "hypocrite" bleed.
>
> Michael Ghiselin, *The Economy of Nature and the Evolution of Sex*

Could morality have evolved? Might the full suite of our ethical responses and judgments, our sentiment of right and wrong, itself be the product of natural selection? A lot of effort has been expended, by biologists and by philosophers, on this question; on attempted demonstrations that ethics might, or did, or could not, have come about by the process of natural selection. At first sight, this is a curious endeavor. After all, what does it matter where ethics comes from – whether it is the product of natural selection, a gift from God, or the reflection of values which are somehow eternally inscribed in the very fabric of the universe (to mention only some of the possibilities). What matters, we might think, is that ethics exists; that our conduct is and ought to be regulated by its demands, and that therefore we ought to get on with clarifying those demands without concerning ourselves with their source. However, there are multiple and intersecting reasons to think that the source of morality has a bearing on how we should behave.

It might be, for example, that a conclusive demonstration that morality could, or could not, be the result of natural selection would strengthen the case for one of the competing theories of the *justification* of our moral code. Some philosophers are skeptics, of one kind or another, about morality. Some are skeptical about its objectivity; for them, the demands and requirements of our morality represent merely one possible form a valid morality might take. Other philosophers are skeptical that its demands upon us are really binding. People who adopt this position might see morality as simply the reflection of class interests, as some Marxists have done, or as the attempt by the weak to ensnare the strong, as Nietzsche suggested. If it could be shown that, as a result of natural selection, we might be expected to share a common morality, certain of these views would seem much more plausible than others. Those philosophers who hold that other, equally valid, moralities are possible would appear to be vindicated. The path taken by evolution was not mapped out; the results of evolution are the contingent product of multiple forces, which might have led elsewhere (and which might yet lead us away from where we find ourselves today). Our moral system would not be uniquely rational, or uniquely moral, as many philosophers have thought. Instead, it would be just one of many, equally valid, possible systems.

To this extent, evolution lends support to philosophical skepticism, but in other ways it tends to undermine it. If morality is the product of evolution, then its demands upon us might be unavoidably binding, rather than the product of mere convention, as some skeptics have thought, and no amount of reflection on the fact (if it is a fact) that it could have been very different will enable us to shake them off. As an example, compare the demands of morality to our perception of color. This perception is also a contingent product of evolution: if evolution had taken a different course, our eyes might be sensitive to a different range of wavelengths of light; perhaps, like those of some insects, they might be attuned to ultraviolet light. If this were the case, then the range of colors we could distinguish would be very different from those with which we are actually familiar. But reflecting on this fact doesn't make our perception of color any less compelling. Similarly, reflecting on the fact that our morality is the contingent product of evolution might do nothing to shake the grip of the emotions that subserve it.

It might also be that debates over the *content* of morality could be settled by evidence about evolution. Those Marxists who see, in conventional morality, a rationalization of class interests will have to confront arguments which show that not merely morality in general, but even some of its fine details (including aspects which they see as expressions of class privilege), are the products of evolution.[36] More centrally, the debate between skeptics about morality (who believe that we are motivated simply by self-interest and we use morality as a convenient cover), and defenders of authentic altruism (who argue that humans are frequently motivated by a genuine desire for the welfare of others), has often been conducted with reference to the evidence from evolution. This debate will serve as a reference point as we explore the evidence that morality is the product of evolution.

What is morality?

Defining morality is no easy task. Fortunately, in everyday contexts, we can to some extent rely on our intuitive grasp of the concept: though we may not be able to articulate our intuitions very clearly, we know morality when we see it. However, if we are going to be able to answer difficult questions about its possible origins, and adjudicate debates between thinkers who argue for an evolutionary source for morality and those who contend that evolution could never give rise to anything more than a simulacrum of the notion, we need to be able to identify at least some of the central planks of morality. We shall need fixed points, to which we can refer and compare the kinds of proto or ersatz moral behaviors yielded by evolutionary models.

To a first approximation, we might describe morality as a system of *prescriptions* that are held to be *unconditionally* binding upon all rational agents. That is, morality is a set of rules, explicit or implicit, which, in so far as they are capable of assessing and controlling their actions, and regardless of their beliefs and desires, each person is required to obey. Morality is not something you can opt out of; it is incumbent upon all autonomous agents. Only those who cannot understand or obey its commands – members of other species, very young children, the insane, and, more controversially, those suffering from impulse control disorders and addictions – are excused from its demands, and then only because they cannot be expected to conform to it, not because it is not true for them. Morality, at least at its core, is objective. It is not open to being altered and its demands are inescapable.

Remember, we are concerned here only with the *concept* of morality. That is, we wish to analyze what we mean when we speak of a moral duty or prohibition. In saying that the concept of morality is of something objective and unconditionally binding, we are not committing ourselves to saying that anything answers to this concept. We are not, for example, taking sides between relativists and absolutists, or between moral realists and anti-realists. Just as we can analyze the concept of God (an eternal, omnipotent and omniscient being) without committing ourselves to being believers, so we can analyze morality without taking sides on moral questions, or even on whether there *are* any moral questions.

The concept of morality I have been expounding thus far, is, roughly, Kantian. That is, it is the notion of morality that received its first full elaboration in the work of the eighteenth century German philosopher, Immanuel Kant, whose work revolutionized all the central fields of philosophy. Kant argued that morality was encapsulated by what he called the *categorical imperative*, which is a rule that is unconditionally binding upon us, as rational agents, and which is delivered to us by our rationality. We are not concerned here with the details of Kant's view, but with its shape: for there can be little doubt that Kant's view of morality as an objective and unconditionally binding system of imperatives captures an important part of our shared concept.

However, it is plausible to think that Kant only gives us half the picture. As many philosophers have pointed out, Kant's morality is rather bloodless and abstract. He argued that an action had moral value only in so far as it was motivated by respect for the moral law alone. But most of us think that moral actions ought to be motivated by concern for other people, not for the moral law. We should, as Michael Stocker famously pointed out, think rather less of the friend who comes to visit us in the hospital because he feels it is a duty incumbent upon him, than of the friend who is motivated by concern for us and our welfare.[37]

Exactly what role desires and emotions ought to play in morality is controversial. But it is less controversial to maintain that they have *some* essential role. We can express the core of this intuition by saying that moral prescriptions are intrinsically *motivating*. There is something very odd about the idea of a person who sincerely assents to the proposition that we ought to give to charity, but isn't motivated actually to do it. To be sure, we are all too well aware that a moral proposition can fail to motivate us *sufficiently* to act upon it: we know, from experience, that we frequently find ourselves backsliding – being stingy with donations to worthy causes, finding excuses for not visiting sick relatives, and so on. Nevertheless, it seems that accepting a moral proposition has to connect up with our motivational system in *some* manner: if not by moving us to act on it, then at least by making us feel shame or guilt at our failure.

Indeed, some philosophers have gone so far as to suggest that there is nothing to morality beyond its subjective side, expressed in feelings, emotions, and dispositions to act. Kant's great rival, across many fields of philosophy, was the Scottish thinker David Hume, and it is with Hume that the idea of morality as, essentially, a set of feelings is most closely associated. By basing morality on feeling, Hume accounts for our conviction that there is an internal connection between accepting the truth of a moral proposition, and being motivated to act on that proposition. But if Kant's moral system seems rather bloodless, then Hume's seems unable to account for the apparent objectivity of morality. Our notion of morality appears to combine both Humean and Kantian ingredients, and any attempt to reduce one to the other captures only part of the concept.

Thus, the concept of morality is of a set of rather strange ("queer," as J. L. Mackie put it) facts: facts that are intrinsically motivating.[38] Most facts are not like that at all: we do not expect people to be moved to action simply by facts about the natural world (though it is common enough to be moved by such facts in conjunction with moral facts, or desires). Mackie thought that such queer facts were incoherent: nothing could answer to such a concept. We have a good grasp on what an objective fact is, a good grasp of what a motivating desire is, and no notion at all about how one and the same thing could be both at once.

The concept of morality, the standard against which we shall measure an evolutionary explanation, is a concept of a set of prescriptions that are objective, universally binding, and intrinsically motivating. Could we have come to possess such a strange concept as a result of evolution? Can the blind forces of natural selection really give rise to such an elaborate intellectual construction? More importantly, could evolution produce beings that can – indeed, are obliged to – *act* morally? Ought we to look elsewhere for the explanation of morality (or, more radically, disabuse ourselves of the notion that we ever do behave morally)?

So far, we have been concerned only with the *form* of morality: what can analysis of our concept of morality tell us about its structure? It is reasonable to believe that conceptual analysis can also tell us a great deal about the *content* of morality. If it does not have the right kind of content, we should be reluctant to call any system of prescriptions a morality, even if it has the formal features we have laid out. A moral system must be devoted, largely if not wholly, to concern for the welfare of other people. To that extent, it stands opposed to selfishness. We fall short of our concept of morality in so far as we act to benefit ourselves, directly or indirectly. Moreover, a morality must systematize norms of justice and fairness: it must prescribe equal treatment for everyone, unless there are relevant differences between them. We ought not to treat others badly unless they have done something to *deserve* such treatment. Morality must disregard arbitrary differences between people, and ignore, perhaps even compensate for, the effects of sheer luck.

If Mackie, with his argument that nothing can correspond to our idea of a moral obligation, presents us with a conceptual challenge to the concept of morality, evolution presents us with an empirical challenge to its content. If we are to explain how morality might have evolved, we not only need to show how we came to have this (allegedly incoherent) idea, but also why its content is precisely the opposite of what we should expect, given what we know about the process of natural selection. Since evolution is the result of a process that systematically favors selfishness, it is difficult to see how it could possibly yield beings who sincerely believe that they ought to be concerned with the welfare of others, or that norms of fairness and justice are binding upon them. Giving an evolutionary explanation of morality seems, at first sight, about as promising a task as giving a theistic explanation for the origins of atheism.

Evolutionary explanations of morality

The prospects for explaining morality as a product of natural selection do not appear to be very good. Think of the slogan often used to describe natural selection: survival of the fittest; this slogan seems to suggest that evolution is a process that rewards selfishness. Recall, briefly, the manner in which natural selection operates: imagine a population of antelopes that is subject to predation by lions. The antelopes have only one means of defense; at the first sign of danger, they flee. Now imagine that, by chance mutation, an antelope is born that is slightly faster than the others, and that the mutation responsible for its greater speed is heritable. On average, this antelope will survive for longer than its slower herdmates, and therefore will, on average, tend to have more offspring. The small statistical difference this advantage represents can, over many generations, be expected to prove decisive: the heritable mutation (let us call it a gene, for simplicity, though we shall have reason to question this common way of talking later) which is responsible for greater speed will gradually become more and more prevalent in the population. Eventually, every antelope in the population can be expected to have a copy of this gene; the gene has reached *fixation.*

Now imagine that a random mutation gives rise to a "helping" behavior. Various kinds of such behaviors can be imagined. Perhaps individuals with the helping gene lag behind, so that slower animals have a smaller chance of being eaten (the danger is now shared; the lion can choose which antelope to devour). Individuals who behave in this manner run a much greater risk than do those who flee. Possibly fatally for them, they will be out-performed, not only by those with the speed gene, but also by those who lack the gene for helping. Since more of them will fall victim to predation, they will have fewer offspring than other members of the group. The gene for helping will rapidly become extinct.

The lesson from this brief review of natural selection seems to be this: evolution rewards selfishness, and nice guys finish last. Even in the absence of predators, conspecifics are in competition with each other. Antelopes compete for the plants they graze; lions compete for prey. Both compete with members of their own species for mates. The competition between members of the same species is more direct and more intense than that between members of different species, since exactly the same set of scarce resources is required by each conspecific. In the struggle for life, it seems, there is no room for sentiment.

Darwin himself noticed this apparent implication of his theory:

> He who was ready to sacrifice his life, as many a savage has been, rather than betray his comrades, would often leave no offspring to inherit his noble nature. The bravest men, who were always willing to come to the front in war, and who freely risked their lives for others, would on an average perish in larger numbers than other men. Therefore it hardly seems probable, that the number of men gifted with such virtues, or that the standard of their excellence, could be increased through natural selection.[39]

Natural selection, it seems, cannot result in the evolution of behavior that is not fundamentally selfish.

Yet apparently altruistic behavior is actually quite common both in human beings and in other animals. One possible explanation of such behavior in humans is that we, and we alone, are able to transcend the limitations of our animal nature. Perhaps, it is suggested, our ancestors behaved selfishly, and therefore prospered (evolutionarily speaking), at the expense of their more altruistic kin. We owe our very existence to that selfishness, since the lineages founded by less selfish animals died out. Nevertheless, we do not have to obey the dictates of our genetic programming. We have suites of selfish instincts that are the result of evolutionary history, but we also have the ability to assess the behaviors these instincts urge upon us.

Though this is a possible explanation of why we have the capacity to act morally, it is not one that we can adopt with much enthusiasm. If this theory is true, at least in the manner in which its proponents have developed it, we should expect genuine morality to be relatively rare. Advocates of this view claim that our strongest emotions are keyed into behavior that benefits us and our genes, which implies that when we act morally, against the dictates of our programming, we do so only at the cost of great effort. The prospects for morality would not be very rosy on this view. Moreover, the proposal smacks of human chauvinism, which many people will find implausible. We are constituted very much like the other primates who are our close relatives; the differences between us are of degree, not of kind. Is it not arrogant to think that though, like them, we are programmed to behave in certain ways, we alone can transcend our program? How do we do this? Are we not biological machines like them? Is not our behavior the result of our evolved capacities? If we are naturally selfish, then perhaps we are deceiving ourselves when we claim that we act morally on *any* occasion.

Moreover, our acts of altruism appear very much like a variety of animal behaviors. If we are willing to defend our group at the cost of our own deaths, then so are bees. If we will selflessly share food with kin, then so will chimpanzees. Many social animals give alarm calls, apparently to warn the other members of the group of the presence of a predator. This looks like altruism: it benefits others while drawing the attention of the predator to the animal giving the call. We have no reason to think that these animals are capable of transcending their genetic programs. If they act altruistically, then altruism must be compatible with natural selection after all. Given these apparent similarities between ourselves and other animals, we should look, in the first place at least, for a unified explanation of our behavior.

What is altruism?

Before we examine this behavior, we need to clarify what we mean by altruism. A behavior is technically altruistic if it benefits others at some cost to the animal whose behavior it is. To express the concepts we need here, it's useful to adopt the so-called "gene's eye view" of evolution. According to this view, the gene, and not the organism or the species, is the unit of selection. That is, evolution is ultimately for the benefit of the genes. As Richard Dawkins, one of the great proponents of this view, puts it, genes build *survival machines* the better to propagate themselves.[40] We tend systematically to overemphasize the importance of these survival machines, the bodies (including the brains) of animals and the morphological characteristics of plants, because they are the kind of entity we are programmed, by our genes, to deal with. But it is the genes which matter, which control the process, and for whose benefit the entire show is run.

This view – let's call it the *Selfish Gene* picture, after Dawkins' famous book – is much misunderstood. How can genes be selfish? After all, they are not conscious entities; they don't have desires or even instincts. Lacking these properties, genes can't literally be selfish, but it is still helpful to think of them *as if* they were. If Dawkins is right, genes are the unit of inheritance, that is, they are the only aspect of our physical constitution that is passed on (in the form of copies) from parents to children (in fact, even Dawkins would admit that this is a simplification, but it is a useful one for our purpose). Genes, unlike bodies or minds, are potentially immortal, in the sense that identical copies can persist for as long as life goes on. Because genes get copied, and bodies don't, bodies are "invisible" to natural selection. Thus, any improvements – or, more likely, impairments – undergone by an organism's body during its lifetime will not be passed on, unless they have a genetic basis. Since genes, and genes alone, get copied and reproduced indefinitely, evolution automatically selects for whatever is in the interests of genes. If a particular genetic mutation arises which causes, in one way or another, that gene to become more numerous in the population, then it will be selected for. It might rapidly go to fixation. Everything happens *as if* genes are selfish, as if evolution is for their benefit, and as if they are pulling our strings.

With this as our background, we are now in a position to define altruism. In the technical sense in which we shall be using it, a behavior is altruistic if, and only if, it increases the *fitness* of other organisms, at some cost to the fitness of the organism whose behavior it is. Fitness is measured in terms of the ability of the organism to propagate its genes (that is, in terms of its ability to reproduce). This allows us to state the problem of altruism quite neatly: why does natural selection not eliminate genes that lead to such altruistic actions? By definition, it seems, such a gene would decrease in frequency in a population, since fewer copies of it would be passed on to the next generation, while copies of the genes contained in other organisms – organisms whose fitness it had enhanced – would increase in frequency. We should expect the altruistic genes to disappear rapidly.

One possible answer to this question I mention only to dismiss. Mightn't it be the case that altruistic genes remain prevalent in the population for the simple reason that there are no alternatives available? If every organism in a population possessed copies of the altruistic gene, then so long as the species survived, so would the gene. Though genes could survive in this manner over the short term, the evolutionary time span with which we are concerned is immensely long – at least three *billion* years. This allows plenty of time for altruistic genes to be eliminated. In every generation there are a number of random genetic mutations, many of which will not code for the observable characteristics of the organism – its *phenotype* – at all. Of those that do influence the phenotype, the majority will be harmful, and will quickly be eliminated. But some will influence the phenotype in such a manner as to cause it to produce copies of itself at an increased rate; these genes will rapidly propagate. Given the frequency of mutations, and the immensity of the evolutionary time span, genes just cannot survive for no other reason than that there is no alternative. Alternatives crop up all the time.

So, how are we to explain the puzzle of altruism? One popular approach is to *explain it away*. According to the biologists who take this view, altruism does not need to be explained, because it does not exist. What they undertake to explain is the *appearance* of altruism. To assess this claim, it will be helpful to have before us some examples of apparently altruistic behavior. Altruistic behavior in animals ranges from the risky to the reckless to the (literally) suicidal:

- Risky behaviors might include the alarm calls given by many animals and birds. When vervet monkeys see a predator, they sound a warning to the rest of the troop, emitting different sounds for different kinds of predators. The troop responds appropriately to these calls: on hearing the "leopard bark," they run up trees; when they hear the "snake call," they stand up on their hind legs and look around them, and so on. This appears to be altruistic behavior on the part of the monkey giving the call because it risks attracting the attention of the predator to itself. It therefore seems likely that monkeys that engage in this behavior will tend, on average, to leave fewer offspring than those that don't. The *troop* benefits from the behavior, but the individual monkey loses.
- Reckless behaviors are those that are apparently *unnecessarily* risky. The monkey that gives a warning call takes a risk, but might make an effort to minimize it. But the animal who engages in reckless behavior seems to run a risk greater than seems to be necessary to

achieve its aims. For instance, the *stotting* of some animals – repeatedly jumping in the air with all four legs straight – seems to be reckless in this sense. Gazelles who spot a wild dog might stott, rather than run away. The stotting serves as an alarm call, but it is a curious one. Not only does it attract the attention of the predator, it also gives it a head start on the chase. It is as though the gazelle is deliberately drawing the fire of the predator: risking its life for the good of the group.

- Suicidal behaviors are those that benefit other members of a group, at the ultimate cost to the organism of its life. A well-known example of such suicidal behavior is the stinging action of the honey bee. A bee which stings an animal threatening its hive dies soon afterwards, because the barbed sting sticks in the animal, and when the bee pulls away its abdomen is ruptured. It seems as though the bee sacrifices its life for the good of the hive.

Darwin suggested *group selection* explained the persistence of such apparent acts of altruism in the animal kingdom. Though individuals who behave altruistically suffer for it, the groups to which they belong do better than those that are not blessed with such unselfish members. Thus, truly altruistic behavior could evolve. Darwin saw in this process the origin of morality:

> It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an increase in the number of well-endowed men and an advancement in the standard of morality will certainly give an immense advantage to one tribe over another. A tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to aid one another, and to sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection.[41]

For most of the twentieth century, biologists frequently invoked a hypothesis something like this to explain acts of apparent altruism. However, the group selection hypothesis faces a major problem. Imagine two groups, one of which has a high proportion of altruists, while the other has none. The altruists engage in risky behavior, with the result that, on average, they leave behind fewer offspring than do selfish individuals. We can measure the fitness of each kind of behavior, and represent it numerically. We'll use "number of offspring" as our measure of fitness. These numbers are for illustration only, but let's suppose that selfish individuals in the group consisting only of selfish individuals have two offspring each, whilst selfish individuals in the group with a high proportion of altruists have, on average, three offspring each, since they are benefitting from the altruistic behavior of their fellows. The altruists, we'll suppose, have 2.5 offspring each: more than the selfish members of the other group, because they, like their selfish fellows, benefit from the altruism of others, but less than the selfish members of their group, due to the risks they run.

Each group has 100 members. Group 1 is 70% altruistic; Group 2 is 100% selfish. Let's examine what happens to each for two generations.

	Group 1	Group 2
First generation:	70A	100S
	30S	
Total:	100	100
Second Generation:	175A (70 x 2.5)	200S (100 x 2)
	90S (30 x 3)	
Total:	265	200
Third Generation:	438A	400S
	270S	
Total:	708	400

Just as Darwin predicted, groups with altruists in them grow more quickly than groups composed exclusively of selfish individuals. Succeeding generations will do even better; if reproduction continued at this rate, in the next generation the altruistic group would contain 1095 altruists, and 810 selfish individuals, while the selfish group would contain only 800 individuals. It was this kind of trend which impressed Darwin and other group selectionists. Though the altruists do less well than the selfish individuals in their group, their group will outperform a wholly selfish group. It looks, therefore, as though altruistic genes could be selected for.

Notice, however, that the *proportion* of altruists in each generation declines. In the first generation, altruists were 70% of the total population. In the second, the proportion has declined to 66%, and by the third it is 61%. In a few generations, the proportion of altruists will drop below 50%, and it will keep declining. One consequence of this fall is that we can expect the benefits of altruism to the group to fall as well. If, for example, the altruistic act is the giving of alarm calls, then the lower the proportion of altruists in the population, the higher the proportion of attacks by predators which are not preceded by a warning (other things being equal). But the smaller the percentage of warnings, the greater the damage inflicted by predators, on altruists and the selfish alike.

Thus, as altruists begin to be outnumbered by selfish individuals, they tend increasingly to pay the costs of their risky behavior without receiving much benefit in turn from other altruists. Their fitness will fall; which translates directly into a fall in their rate of increase. Of course, the rate of increase of selfish individuals will fall too. The mixed group will continue to reproduce more quickly than the wholly selfish group – indeed, it may well supplant it entirely. But as the proportion of altruists falls, the mixed group begins more and more to resemble the selfish group. Eventually, the proportion of altruists will dwindle to zero. Altruism will have allowed the group to out-compete its rivals; on this point the group selectionists are correct. But so long as the process continues unchecked, altruism is destined to disappear. Ultimately, the formerly altruistic group will be just as selfish as the group it has supplanted.

Richard Dawkins puts the point in this way: altruistic groups might out-compete selfish groups, but they are vulnerable to *subversion from within.* Because, within any group, altruists are outperformed by the selfish, they will eventually be driven to extinction. Ever since George Williams drew this melancholy conclusion to the attention of biologists, group selection has been widely regarded as a dead letter.

However, those who drew the conclusion that altruism could not evolve by group selection did so too hastily. Group selection is not *impossible*; it just requires a very special set of circumstances. We saw that so long as the group continues in existence, the proportion of altruists in it declines, and as a consequence the benefits they bring, to each other and to selfish individuals, fall as well. Altruism can evolve by group selection just so long as groups do *not* stay together. Instead, the group must establish colonies, and the proportion of altruists in the "daughter" colonies must be higher than the proportion of altruists in the mother group.[42] It is certainly possible that both these conditions can be satisfied. Our figures showed that the absolute number of altruists in the group could be expected to increase, at least initially. We also noted that the altruistic group could out-compete selfish groups. If the altruistic group were to take advantage of the demise of the selfish group by colonizing its territory, and if the colonizing populations themselves consist disproportionately of altruists, then the new group can reap the benefits of altruism. Of course, the proportion of altruists in the colony is destined to decline, eventually necessitating a new round of colonization. So long as altruistic groups produce offshoots at a great enough rate, and a high enough proportion of the members of the new colonies are altruistic, altruism can prosper via group selection. Interestingly, if this occurs, then some of the groups that are out-competed by altruistic colonies might themselves be formerly altruistic mother populations.

How frequently is this special set of circumstances encountered in nature? Some biologists believe that it is common enough to be the source of at least some of the altruism we observe in other species, and perhaps of human altruism as well. However, it is clear that group selection can only be part of the story. Though many animal groups splinter in various ways, often as a result of young adult males leaving, or being driven out of, the group, there is little evidence that new populations differ in any significant way from their parent group, so far as the proportion of altruists is concerned. Nor is there any evidence of such selective colonization by early humans in the EEA, the environment in which, most evolutionary psychologists believe, our predispositions to behavior were laid down. The likelihood is that if group selection, as we have outlined it so far, is the only source of altruism, then altruism will be a rare commodity.

As we saw, Darwin, and after him many evolutionary biologists, believed that group selection accounted for altruism. But it is the apparent failure of group selectionist hypotheses, which, more than anything else, explains the current fashion among biologists for dismissing altruism. Faced with instances of apparent altruism, they seek to explain them away, to show that what seems to be altruism is really disguised selfishness. Darwin was right, these biologists believe: only group selection could explain the existence of altruism. But group selection is rare or non-existent, and therefore altruism is absent from the natural world.

I divided apparent acts of altruism into three categories, each one apparently harder to interpret as disguised selfishness than the last. Can merely risky behavior really be selfish? Biologists have expended a great deal of ingenuity on explaining these phenomena. As we saw, giving an alarm call is a risky behavior, which is to say that it is, apparently, at least minimally altruistic. It seems that animals that give such calls lower their own fitness while raising the fitness of others. Biologists who seek to explain away apparent altruism therefore owe us an explanation which demonstrates that this behavior does not lower the fitness of those who practice it.

In fact, there are a number of credible explanations. I'll briefly sketch two, both due to Richard Dawkins. He calls the first the *cave* theory, from the Latin word used by English schoolboys to warn of an approaching teacher. Dawkins' idea is that these schoolboys do not act for the sake of others, but for their own sake. They warn one another of the teacher's approach because they know that they are more likely to avoid strife *themselves* if their peers also refrain from further misbehavior. Similarly, Dawkins conjectures, animals that give alarm calls do so not to help each other, or their group, but to ensure their own safety.

It is easy to see how this might work. A flock of birds is rooting in the undergrowth for food. One of them happens to glance up and notices an eagle circling. The predator has not yet spotted the group, but it is only a matter of time. How should this particular animal act – what kind of behavior will evolution select for? The creature might freeze, hoping that another member of the group might catch the eagle's attention. This would be straightforwardly selfish behavior. But, if it adopts this course of action, it has not done everything it can to minimize its own risk. The longer its fellows continue to move about openly and noisily, the greater the chance the eagle will see the group, and, once its attention is drawn, it just might spot the selfish animal first. Wouldn't the bird do better to hiss a warning, directing the whole group to freeze into immobility, and thus reduce the chances of the eagle spotting any member of the group? Mightn't this course of action be the best, measured in the selfish terms of reproductive fitness?

Dawkins' second explanation is also an attempt to show that giving the alarm is selfish. He calls it the "never break ranks" theory. Imagine a flock of birds, one of which spots an eagle as before. In this case, however, freezing is not as a good an option as fleeing into the branches of a nearby tree, where the foliage is too thick for an eagle to follow. A selfish bird might take to the wing immediately, attempting to get to safety before any of its flock-mates have even noticed the predator. However, as soon as it takes to the air it risks drawing attention to itself; it will be a lone target horribly exposed to the swooping eagle. Far better to give a loud call, causing the entire flock to rise and make for the trees at once. If it follows this course, it will be one bird among hundreds, perhaps even thousands, and its personal risk of falling victim to the eagle will be very small. Thus, the apparently altruistic act of giving an alarm call is shown to be an entirely selfish action after all; an action adopted because it minimizes the risk to the caller.[43]

We might find this strategy plausible when it is applied to risky behavior, but surely reckless actions cannot be explained in the same way? Surely the stotting gazelle's actions cannot be shown to be selfish? In fact, most biologists believe that stotting is not altruistic. Indeed, it is not directed at the gazelle's fellow herd members at all; any benefit they receive from it, such as being alerted to the presence of a predator, is incidental to its function. Rather, stotting is directed at the predator, and its function is to demonstrate the stotting animal's health and vigor. Gazelles who leap athletically thereby advertise that they would be difficult to hunt down. Their display therefore encourages predators to look elsewhere for their next meal: to other members of the herd. Far from being an altruistic act, stotting is selfish. It evolved because gazelles that stott survive at the expense of those who don't.

Kin selection

We might find these reinterpretations of apparently altruistic risky and reckless behaviors reasonable, but find it hard to apply such a strategy to suicidal actions. How can these actions be in the best interests of the organisms who engage in them? Bees that sting animals that threaten the hive do not simply *seem* to endanger their own lives; they (almost always) *actually* die. It seems as though these bees sacrifice their lives for a greater good, for the hive as a whole. Surely this, at least, is a genuinely altruistic act?

It is in explaining this kind of action that the "gene's eye view" really comes into its own. From this perspective, you will recall, everything happens as if evolution is for the benefit of the genes. Genes which, in whatever manner, contribute to behaviors that lead to an increase in their numbers in the next generation will be selected for by evolution. Dawkins and other gene-selectionists often say that genes, unlike the bodies they help make and which carry them, are potentially immortal. Strictly speaking, this is false, for a reason that is important here. Genes do not usually *themselves* outlive the bodies that are their vehicles. Instead, it is *copies* of genes that live on in the next generation. Reproductively successful organisms leave a greater representation of their genes in the next generation, in the form of such copies, than do the less successful.

Once we realize that it is copies that count, we see that organisms have a range of strategies available to maximize reproductive success. The most obvious is the most direct: have offspring. Genes are arranged on chromosomes, long strands of DNA. There are forty-six such strands in humans, arranged in twenty-three pairs. Organisms that have this pairing arrangement are known as *diploid*. Each of our cells is diploid, which is to say that they have two sets of twenty-three chromosomes. There is one exception to this rule, however: sex cells (sperm in males and ova in females). These cells are *haploid*; they contain only a single set of twenty-three chromosomes. In sexual reproduction, the haploid cells of males combine with those of females to create a new organism which, like its parents, is diploid: it receives a single set of chromosomes from each parent.

Each of our sex cells has a (more or less) random selection of our genes within it. Since we are diploid animals, and our sex cells are haploid, each gene has a 50% chance of finding its way into each sex cell (though, as we might expect, natural selection favors genes which are able to distort this process in their favor).[44] The upshot of all this is that for any particular gene, the chances that a single offspring of an animal will possess a copy of it by descent are 50%. In other words, the degree of relatedness between parents and offspring in a sexually reproducing species is 50%.

This "coefficient of relatedness" (to use the jargon of biology), is exactly the same as that which holds between full siblings in sexually reproducing diploid species. For each gene you receive from your mother, there is a 50% chance that you share it with your sister or brother (since you each got a random selection of 50% of her genes); that is, your degree of relatedness on your mother's side is 25% (50/2) or 0.25. But if you also share the same father, then you have precisely the same degree of relatedness on his side, giving a total degree of relatedness of 50% or 0.5. We can use the same logic to show that grandparents and grandchildren have a coefficient of relatedness of 0.25, as do half-siblings, and uncles and aunts with their nieces and nephews. First cousins have a coefficient of relatedness of 0.125, as do great-grandparents and their great-grandchildren. And so on.

Since reproductive success, measured from the gene's eye view, concerns the extent to which we are able to get *copies* of our genes into the next generation, we can increase our fitness in one of two ways: either by having offspring of our own or by taking steps to ensure the reproductive success of our close relatives. Frequently, of course, we can do both, but sometimes we have to choose; when we are faced with such a choice, the best course of action (again, from the gene's eye view) will depend on the circumstances. If my circumstances are such that I cannot afford to raise offspring of my own – my resources are limited in some way, or my chances of securing a mate are low – I might do best by assisting others. Even if I am able to have offspring of my own, it might be that I do better by refraining. Imagine a case in which a diploid, sexually reproducing animal is faced with a choice between bearing one offspring itself, and assisting a full sibling to raise three. Since the coefficient of relatedness between such an animal and its own offspring is 0.5, it does better by raising three nephews and nieces. Each such relative has a degree of relatedness to it of 0.25, and 0.25 multiplied by three is greater than 0.5.

Biologists use the term *inclusive fitness* to refer to reproductive success in this extended sense. The inclusive fitness of an organism is a measure of its success in increasing the proportion of copies of its genes in the next generation, by whatever means. From the perspective of inclusive fitness, many apparently altruistic acts can be seen to be instances of genetic selfishness. We can, for example, see why it might be sensible, from a genetic point of view, for an organism to forgo having offspring of its own, in order to assist its kin with raising theirs. We can even see how, under the right circumstances, it might make sense for an organism to sacrifice its life for its relatives. J. B. S. Haldane, the great evolutionary biologist, was reportedly asked if he would lay down his life for a brother. "No," he replied, "not for fewer than two brothers, or eight first cousins."

Biologists call the process of natural selection through enhancement of the reproductive fitness of close relatives *kin selection.* Kin selection allows us to explain the existence of many kinds of behavior which otherwise seem quite mysterious. A puzzle remains, however. Though it is apparent that reckless behavior, and even, in the right circumstance, suicidal behavior, can be selected for through kin selection, how do we account for the fact that bees (and other social insects) so willingly and frequently lay down their lives for the community? Though it may be understandable that diploid, sexually reproducing beings, like ourselves, would sacrifice their lives for their kin *in extremis*, how do we explain that bees' very first line of defense involves the ultimate sacrifice?

The clue lies in the unique system of reproduction of most of the social insects. To take bees as an example: almost all the bees in a hive are sterile; their reproductive systems shut down by the pheromones released by the queen, who is the mother of the entire hive. Her female offspring, who become the workers and guards of the hive, are all full sisters. But, due to an oddity of their reproductive system, they are more closely related to one another than are human siblings. The Hymenoptera – ants, bees, and wasps – are haplo-diploid. Female bees are diploid: they hatch from fertilized eggs and so possess two sets of chromosomes, one from each parent. But male bees hatch out of unfertilized eggs, and so have only one parent: their mother. Thus, no male has a father, or any sons. As a result, male bees are haploid, possessing only a single set of chromosomes. In consequence, every female bee has a 50% chance of sharing any one of her mother's genes, but, since her father had only one set of chromosomes to pass on, she has all her father's genes, as do all her female siblings. Full sisters therefore have a coefficient of relatedness of 0.75, rather than 0.5. So, they are more closely related to their sisters than they would be to offspring of their own, should they have any! This fact explains why bees are better off sterile, assisting their queen to produce more near-clones of themselves, than they would be rebelling and going in for reproducing themselves. It also explains why a propensity to lay down their lives for one another has been selected for. Such "sacrifices" are simply one more way in which a bee acts – selfishly – to ensure that copies of its genes will be represented in the next generation.[45]

Reciprocal altruism and game theory

I have sketched a variety of mechanisms whereby altruism is shown to be disguised selfishness. Apparently altruistic acts may be aimed quite directly at the good of the organism, as in the case of alarm calls and stotting, or they might be instances of kin selection, in which an individual sacrifices his or her own reproductive interests for the sake of their investment in the reproduction of close kin. It is unlikely, however, that all apparently altruistic acts can be explained through these mechanisms. Not all altruistic acts are aimed at close kin, and not all are amenable to interpretation as direct (though disguised) selfishness. Sometimes, animals simply seem to help one another. Indeed, sometimes this helping behavior crosses species boundaries.

Over the last twenty years, evolutionary biologists have turned to the mathematical discipline of *game theory* to aid them in understanding these phenomena. Game theory provides us with a set of tools with which to model *strategic* interaction. A "game" – here the term is used to refer to any kind of interaction between two or more players in which there is a question of winning and losing, profit and loss – is strategic if the best "move" depends not merely on the state of the game, but also upon what the other players do. Most ordinary games are strategic in this sense: the best spot to place a lob in tennis depends crucially on where one's opponent is moving. Thus cricket, football, chess and poker are all strategic games, whereas golf, in which the player competes against the course as much as opponents, is not importantly strategic.

Most of the games we are familiar with are *zero-sum* games. In a zero sum game, the gains of one player automatically translate into the losses of another. In these games, cooperation between opponents is out of the question: only one player or team can win. But the games which interest us most here are *non-zero-sum* games, in which it is at least possible for separate players to do well without their gains coming at the expense of others' losses. Game theorists are especially interested in games which may appear to their players to be zero-sum games, but which, from the appropriate viewpoint, can be seen to be non-zero-sum. Many economic interactions are like this: participants see themselves as competing with one another for scarce resources, but if they cooperate with one another, they might increase their returns.

The most famous game of all is known as the *prisoner's dilemma*:

> Two prisoners are being interrogated separately by the police. They are accused of committing a crime together, but the police do not have sufficient evidence to convict them. Each is offered the same deal: if they will confess their guilt, but agree to testify against their codefendant, they will be released on a good behavior bond. If they stay silent, however, and their codefendant accepts the deal, it will be she (or he) who is released, while they go to jail for ten years. If both the accused confess, each will go to jail for five years. And if neither confesses, the police will be unable to secure a conviction. However, they will be charged with and convicted of with some lesser crime – perhaps resisting arrest – and each will receive a six month jail sentence.

Game theorists construct a *pay-off matrix* to model this kind of situation. Let's call staying silent *cooperate*, and confessing *defect*. In this matrix, the options for player one are displayed to the left of the boxes, while those for player two are above the boxes. The top set of numbers in each box represents the pay-off to each player, in years, with the pay-off to the first player separated from that to the second by a comma. The second set of numbers, in brackets, represents each player's ranking of the options, from most preferred (1) to least preferred (4).

	Cooperate	Defect
Cooperate	1/2, 1/2 (2,2)	10, 0 (4,1)
Defect	0,10 (1,4)	5,5 (3,3)

Thus, player one would most prefer to defect while player two cooperates. If she were able to secure this result, she would go free, while player two would go to jail for ten years (we assume that each player is concerned with only their own welfare, which is to say minimizing their jail-time, and does not care one way or the other what happens to the other player). If player one cannot secure this result, then she would most like her second-ranked preference, in which both players cooperate, to be the outcome, since in this scenario she will go to jail for just six months. The situation in which both players defect is ranked third by her (five years' jail each), while the outcome she most wants to avoid is that in which she cooperates while her co-accused confesses, since in this situation she would end up serving ten years in jail.

We can see immediately that this is a non-zero-sum game. Though it is possible for one player to "win" at the expense of the other – if one cooperates while the other defects – it is also possible, if both cooperate, for both to secure a good result. And since it is extremely unlikely that either would be able to convince the other to cooperate while they defect – remember, we are assuming self-interest here – it seems obvious that the outcome they should strive to bring about is mutual cooperation. However, while they might both prefer this result to any other (with the exception of the unobtainable situation in which they defect on a cooperator), and it would be a mutually satisfying arrangement, it is not clear that it is available to the players.

To see this, we need to realize that each player is better off defecting, no matter what the other player does. If Jack cooperates, then Jill is better off defecting – she gets off scot-free, rather than going to prison for six months. But if Jack defects, then Jill had better defect as well, since if she cooperates with a defector, she goes to jail for five years. In the jargon of game-theory, "defect" is the *dominant* strategy, which is to say that it ought to be chosen no matter what strategy is employed by the other player. However, the prisoner's dilemma is a symmetrical game: whatever holds for one of the players is automatically true for the other. If "defect" is dominant for Jill, then we may be sure that it is dominant for Jack (we can easily confirm this by noticing that no matter what Jill does, Jack is better off defecting). Since "defect" is dominant for both players, that is just what they will do, if they are rational. As a result, each will go to prison for five years.

But if mutual defection guarantees the players a worse result than mutual cooperation, than cooperation is in both players' interests. If this is the case, surely, it is rational for them to cooperate, not defect. There must, we cannot help but think, be some way in which the players can come to an agreement, and secure a better outcome for each. But how? An obvious way to proceed is via explicit bargaining between the players. Perhaps the unsatisfactory outcome of the prisoner's dilemma is the result of the isolation of the players. The police, we might think, separate them in order to prevent them from coming to an agreement. If we allow them to discuss the situation, we might hope to secure mutual cooperation.

Very well then, let's try it. Having been interrogated separately and informed of their choices, Jack and Jill are sent back to adjoining cells to think things over. There, they discuss their predicament. Each sees that mutual cooperation is preferable to mutual defection, so they come to an agreement. They promise each other that when they are interrogated again, they will cooperate; in other words, they will not testify against each other. Now, if each keeps his or her side of the bargain, they will avoid five years in prison each. But what ensures that they will stick to their agreement? As Jill is taken back to the interview room, she might reason thus:

> Suppose Jack sticks to the agreement. Then he will cooperate. In that case, if I cooperate I will go to prison for only six months. But if I defect then I won't go to prison at all. If Jack cooperates, then I'm better off defecting. But if I'm better off defecting, then so is Jack. Surely he'll see this, and defect. And if he defects, I had better do the same, to avoid ten years in prison. So whatever Jack does, I ought to defect.

Of course, Jack will defect as well. So both players will go to jail for five years each. It is difficult to see how this result can be avoided. Though mutual cooperation is in each player's best interests, mutual defection seems the inevitable result.

What has all this to do with evolution? At first sight, very little. One way to express the paradoxical implications of prisoner's dilemma-type situations is to say that when its conditions apply, rational agents do less well than irrational. Whereas two irrational agents might cooperate because they fail to see that "defect" is the dominant strategy, rational individuals will recognize that whatever the other player does, they are better off defecting, will act accordingly, and end up worse off than irrational agents. Though this is certainly a fascinating discovery, it seems quite irrelevant to our question. We are concerned, here, with whether morality might have evolved. To approach this question, I have focused mainly on non-rational animals, from bees to birds. I have taken this approach because I am concerned with discovering what kinds of dispositions and motivations our evolutionary history might have bequeathed to us, on the assumption that our fundamental desires will, to some extent, reflect that history. Now we discover that in certain kinds of situation it is difficult or impossible to bring rational agents to cooperate with one another, even though doing so is in their interest. This might be bad news to some, but, we might think, neutral or even good news for us. Since we are concerned with what kinds of dispositions we might have developed *before* (or perhaps at the same time as) we evolved rationality, the fact that rationality can be a barrier to cooperation is irrelevant.

Unfortunately, the apparently tragic implications of the prisoner's dilemma for human cooperation cannot be so easily evaded by the evolutionist. It is certainly true that our distant ancestors did not and could not reason as to what strategy to utilize when they found themselves confronted with a choice between cooperation and defection. We, like every other living thing on this planet, are descended from the one-celled organisms that were the first living creatures; such organisms had no capacity for any kind of thought at all. But even single-celled organisms like bacteria can find themselves in prisoner's dilemma-type situations, and the reasoning which shows that humans will engage in mutual defection in such circumstances seems also to apply to bacteria.

Imagine, for instance, a group of bacteria, of a single species, occupying a small rock pool. The number of bacteria that the pool can maintain is strictly limited, because each individual produces waste products, which break down only slowly. These waste products pollute the bacteria's environment; if the pollution exceeds a certain level, a mass die-off will occur. It might even be that death on that scale will cause further pollution of the water, leading to the extinction of the entire population. So, it is clearly in the interests of the bacteria as a group to limit their population. If they can maintain their numbers below a critical threshold, which I shall call *n*, then the population will thrive, but if they exceed it, then the entire group faces extinction.

This scenario can easily be modeled using game theory. Here, "cooperate" translates as "limit your rate of reproduction, so as to cause no net increase in numbers," and "defect" as "reproduce at some (unspecified) faster rate". Imagine, further, that all bacteria currently follow the strategy of cooperating. What does it mean for a bacterium to implement a strategy? Clearly, it cannot mean that it weighs up the consequences of various alternative actions, and selects that action which has the best outcome. A bacterium is not capable of sophisticated mental processing; indeed, it is not capable of mental processing at all. All we mean by saying that it follows a strategy is that it tends to act in some manner that is in accordance with one of the alternatives in our model. In this sense, bees, when they lay down their lives for the hive, are following a strategy that we might usefully label "cooperate." Bees do not deliberate about how to act; they follow a program that has been laid down for them by natural selection. Bees which behaved in this manner in the past had more descendants (better, possessed genes which increased in number in future generations) than did those which did not, and therefore the genes which encode this manner of behaving gradually went to fixation in the population, which is to say that all bees now behave in this manner when faced with those circumstances.

If bees follow a program, this is even truer of bacteria. We assume, therefore, that all the bacteria in our population follow the programmed strategy of cooperating, and the population in the rock pool remains below *n*. But now suppose that a mutation occurs among the bacteria, so that individuals with that mutation will defect (that is, reproduce at a faster rate). How likely is such a mutation? Given the length of evolutionary time, the rate at which mutations occur, and the short generation time of bacteria, it is extremely likely, as long as the behavior for which it codes is not very different from those in which the bacteria already engages, and it has no special costs in terms of the resources required to sustain it. (It is extremely unlikely, for example, that a mutation will suddenly occur amongst our bacteria which enables those who inherit it to fly, since flying is an ability that requires a great many evolutionary steps.) Since the mutation in question seems to meet these conditions, it is likely that it will occur, sooner or later.

Bacteria with the "defect" mutation will, by definition, tend to leave more offspring than those without. For this reason alone, they are fitter than bacteria that cooperate. We can therefore expect the "defect" mutation rapidly to go to fixation. Of course, when this occurs the number of bacteria in the population rapidly exceeds *n*, with the result that the whole population goes extinct! Even so, the result might be inevitable, for this reason: each individual bacterium is better off defecting than it is cooperating. If all the other bacteria cooperate, then a single defector increases its fitness at no cost, because it is unlikely that the defection of any one bacterium will causes the population to exceed *n*. But if the other bacteria defect, then it will inevitably pay the cost that results from increasing pollution, and therefore must seek to produce as many copies of its genes as possible, in the hope that one will manage to survive the coming cataclysm. In other words, "defect" is dominant for the bacteria, just as it was for the rational players of the prisoner's dilemma.

This is bad news for us, for it seems to indicate that cooperation cannot evolve by natural selection. Since cooperation is an important part of morality – especially when the alternative is selfishly doing someone else down to get some benefit – the apparent failure to demonstrate its compatibility with natural selection seems to reinforce the case of those people who hold that evolution is fundamentally unethical (which presents us with a further dilemma: either we accept the truth of evolution, and give up on ethics, or we attempt to find grounds for rejecting, or at least limiting the power of, evolution in the name of ethics).

The prisoner's dilemma iterated

Our failure to "solve" the prisoner's dilemma seems, in one important sense, to be bad news for those who wish to vindicate morality. It apparently suggests that skeptics or cynics about morality are right; morality might be no more than disguised self-interest. We have reason to sacrifice our own narrow interests for those of our close genetic relatives, but these reasons are ultimately self-serving; they represent the victory of genetic selfishness, rather than of moral selflessness. We sacrifice ourselves for others, but only because we (rightly, from a genetic point of view) regard them as, in some sense, extensions of ourselves. If we are right in thinking that morality requires us to give at least some weight to everyone's interests, regardless of the degree to which they are related to us, then it seems it cannot evolve.

We are forced to this melancholy conclusion on the condition that the kind of model for strategic interaction we have just constructed captures the real-world interaction of organisms accurately. However, we have good reason to think that a great deal of the interaction between potential game players is significantly different from the prisoner's dilemma model in one important respect: we deliberately structured our game so that each player chose only once. Though there are many such one round (in game theory, *one shot*) games, many others are *iterated.* Iterate the prisoner's dilemma, and the incentives for cooperation are greatly increased.

Of course, a genuine prisoner's dilemma – in which the pay-offs are jail sentences – cannot be an iterated game. In any possible outcome of a single round of such a game, at least one player is in prison, and therefore no longer available as a strategic partner. So, let us change the scenario to make the iterated game a possibility. For the moment, we won't attempt to model a real-life situation, but will be content with an abstract model. Imagine a version of prisoner's dilemma, then, in which two players compete for money. Obviously, this kind of game can be repeated indefinitely. To fix our ideas, we can assign the following values to the pay-offs: the *temptation*, the amount of money a player receives if they defect while the other player cooperates, will be $8; the *reward*, the money each player receives for cooperating with a cooperator will be $5; the *punishment* for mutual defection $2 and the *sucker's pay-off*, received by the unfortunate player who cooperates with a defector, will be zero. Once again, "defect" is the dominant strategy: in any particular round each player is better of defecting, no matter what the other player does. In a one shot game with this structure, rational players will defect. But what is true of one shot games is not necessarily true of iterated games.

Robert Axelrod, an American political scientist, set out to discover the best strategy to follow in an iterated prisoner's dilemma. He utilized a novel method to test various strategies: he ran a tournament, and invited game theorists, political scientists and psychologists to submit strategies. Axelrod then ran the strategies against each other on a computer. Each strategy competed against every other strategy, and itself, in iterated games of prisoner's dilemma. A variety of strategies were submitted, some of them very complex and subtle. But the winner of the tournament was one of the simplest strategies: a strategy now known as tit-for-tat (TFT). TFT began each game against a new player by cooperating; thereafter, it simply copied whatever move the other player made last time. Thus, if the second player cooperated, TFT cooperated as well, and did not defect unless the other player defected first. In the technical vocabulary of game theory, this makes TFT a *nice* strategy. It is also a *forgiving* strategy, in that it only punishes defectors once. If the other player returns to cooperating, TFT responds in kind.

Why does TFT out-perform other, less nice, strategies, when "defect" is the dominant strategy? Though each player will do better, in any *particular* round, by defecting rather than cooperating, when their scores are tallied over many rounds, players who always defect do badly. We can see this by considering the iterated prisoner's dilemma in an evolutionary context. Earlier, we saw how a population of organisms which adopted the strategy "always cooperate" was vulnerable to invasion and eventual displacement by mutants playing "always defect" (subversion from within). But now let's examine what happens when we introduce a third strategy into the mix: TFT.

Richard Dawkins has provided a thought experiment to model such a situation. He asks us to imagine a population of birds, who are parasitized by a tick. The tick must be removed, because it carries a fatal disease, which the birds will contract if the tick is left too long. Each bird can remove ticks from its own body, but it cannot reach the back of its own head, and so requires the cooperation of another bird to remove ticks from there. Dawkins supposes that the population is initially composed of cooperators. But, he points out, such a population is vulnerable to invasion by defectors, who accept grooming from any other bird, but never groom others in return. Since grooming other birds has a cost (in terms of time which could have been used for foraging), birds that never groom will do better than those that do. Thus, defectors will be slightly fitter than cooperators, with the result that this strategy will spread. Eventually, "defect" will go to fixation.

But now imagine that, by chance mutation a third strategy arises. These birds play TFT: they willingly groom any bird once, but if that bird fails to reciprocate, they refuse it further grooming. How would such a mutant fare in a population of defectors? A lone TFTer would do very badly: it would spend a lot of time grooming other birds – since it will groom any bird once – and will never be groomed in return. Its sad fate will simply be to contribute slightly to the fitness of the defectors, perhaps before dying of the tick-borne disease. But if TFT can get a foothold, however small, in the population, everything changes. Two TFTers, who associate preferentially with one another, can do better than the population average, since they are assured of having their ticks removed, whereas defectors can only have their ticks removed once each. So long as the benefit they receive from being groomed by each other exceeds the cost of grooming, they will be fitter than the defectors. If this is the case, then TFT may very well increase in the population.

Thus, a population of cooperators is susceptible to invasion by defectors, and TFTers can invade a population of defectors. But TFT will not go to fixation. As the proportion of TFTers increases in the population, the probability that any given defector will meet a TFTer with which it has not previously interacted increases. Thus, its chances of being groomed rise, and a small number of defectors may be able to hold on. In addition, a large number of TFTers provide a hospitable environment for cooperators. Indeed, when cooperators interact with TFTers, they are indistinguishable from each other. As a result, the number of cooperators in the population may rise, which in turn provides opportunities for defectors.

Because no one strategy has a decisive advantage over the others, none will go to fixation. What actually happens will depend on the details of the situation – exactly how much it costs to groom other birds, or the penalty incurred by failure to be groomed – but the upshot will be a *polymorphism* of strategies, in which each strategy is represented in some proportion. Any departures from this polymorphism increase the pay-offs to other strategies, which leads to an increase in these strategies, which reduces the pay-offs to the first strategy. For example, if the number of cooperators increases, defection becomes more profitable, as defectors are able to take advantage of cooperators. As a result, cooperation becomes a less profitable strategy, and the number of cooperators falls, leading to a decrease in the number of defectors who prey on them. So a polymorphism of this kind is not static, but it is quite stable over time. Unless conditions change, so that the pay-offs alter, departures from such a polymorphism are usually small and brief.

There are three ways such a polymorphism could be realized. One is in the manner we have assumed in the foregoing, in which different individuals reliably play different strategies. Or a polymorphism could be realized by individuals playing all of the strategies with a certain probability: that is, the same individual sometimes behaves like a cooperator, sometimes like a TFTer, and sometimes like a defector. Or, finally, a population might consist of a combination of all kinds of individuals, in which some always cooperate, some always defect, some always play TFT, and some switch between strategies.

Perhaps the world we live in and share with others is just such a stable polymorphism, in which most people cooperate, or play tit-for-tat, most of the time, but where some people always cooperate (we call them saints), some always defect (we call them evil), and others switch between strategies depending on their costs and benefits. None of us will be surprised to learn that usually reliable cooperators become defectors when the risk of detection is low. But, though it is easy to use game theory to model the behavior of rational actors, and the results may be quite plausible as a description of actual human society, can it be applied to the behavior of other organisms, which do not possess our ability to predict the consequences of behavior? We have seen that, when we limit ourselves to simple strategies such as "always cooperate" and "always defect," rationality is not required for organisms to pursue strategies, nor for processes to occur which can be described using the tools of game theory. But tit-for-tat, surely, is beyond the cognitive grasp of all but the most complex animals – to say nothing of the complex accounting required when individuals switch between strategies. Since behavior in the "lower" animals is programmed by their genes, we can predict that the strategies they pursue will be simple. But that is bad news. As we are concerned with whether morality might have evolved, if conditional strategies such as TFT are beyond the capability of most or all non-human animals, and we must therefore limit ourselves to "cooperate" and "defect," then we can predict that morality cannot have evolved. Populations of cooperators are vulnerable to invasion, which implies that morality must be a recent innovation, if indeed it exists at all.

Fortunately for us, there is evidence that strategies like TFT do not require rational minds to implement them. Cooperation based on upon reciprocity is a genuine feature of the animal world. The most famous example here is a slightly macabre one: the behavior of vampire bats. Vampire bats, as their name suggests, live on the blood of other animals. Each night they seek a large mammal, and try to inflict a painless bite upon it. They then drink blood for around thirty minutes before flying back to their roost. The bats need to feed almost every night: if the hunt fails two nights in a row, they risk starvation. On any given night, from seven to thirty per cent of the bats in a colony will fail in their search for blood. However, those that are successful are able to store blood in their stomachs, and regurgitate it. They can therefore donate this blood to other bats. Gerald Wilkinson studied the bats for five years, and discovered that the benefits of such donations to starving recipients exceeded the cost to donors.[46] A blood donation was worth about eighteen extra hours of life to a starving vampire bat: sufficient time to hunt again. But the donation cost only about six hours of the time a satiated bat had until starvation. In this situation, all the ingredients for a prisoner’s dilemma are in place. We have the following pay-off matrix:

	Cooperate (Feed)	Defect (Don’t Feed)
Cooperate	12,12 (2,2)	–6,18 (4,1)
Defect	18,-6 (1,4)	0,0 (3,3)

The benefit of cooperation is measured at twelve hours of life, which is the eighteen hours gained by a starving bat that is fed, minus the six hours’ worth of life the feeding bat expends in cooperating. We can see that the preference ordering for each bat (once again represented by the numbers in brackets) is identical to that in the classic prisoner’s dilemma. This means that “defect” is dominant: no matter what the other bat does, each bat is better off if it defects. Yet bats do frequently feed one another. Given that this is the case, we can predict that the three conditions necessary for the evolution of TFT must be fulfilled in vampire bat colonies: the game is iterated, the bats interact with one another repeatedly, and they are able to recognize one another.

It is easy enough to see why the game is iterated. Night after night the bats fly out of their cave or hollow tree in search of blood. The bat that is successful on one night might fail the next; thus each bat will frequently find itself in a position to play one or other of the roles in the game. Wilkinson found, moreover, that bats tend to roost together. Vampire bats live in groups that share a roosting spot; though one will occasionally leave for another group, in general the groups are very stable. Thus, the same bats interact with one another repeatedly. Finally, Wilkinson found that bats could recognize each other. In these circumstances, it is not surprising that bats do not follow the strategy that is dominant in the one shot game. Instead, they play TFT. Wilkinson tested this hypothesis by removing bats from different colonies, and starving one at random for a single night. He found that the bats were much more likely to regurgitate blood to the hungry bat when it was returned to their cage if it was a bat that had previously fed them. Thus, it seems that bats keep some kind of rough score, good enough to be able to play TFT. TFT does not, it seems, require the sophisticated cognitive equipment of a primate.

Indeed, there is evidence that much simpler animals than even vampire bats can play TFT. Robert Trivers, one of the most important figures in the development of models to study reciprocal altruism in non-human animals, suggests that TFT explains the relationship between predatory fish and the much smaller fish that clean them. These fish are often the right size to make a good meal for the predators. Yet they do not attempt to eat the cleaners; instead, they seek them out, and when they locate them, go into a kind of trance while the smaller fish removes parasites from their bodies. The small fish sometimes actually swim into the mouth of the larger, and out of their gills, in the search for the parasites.

It is obvious that each player in this game stands to benefit from the interaction. The large fish have their parasites, which otherwise might cause serious problems, removed. The small fish gain a meal. But why don't the bigger fish defect? Why don't these fish accept the cleaning services of the smaller, and then round off the experience by eating them? The answer seems to lie in the fact that cleaner fish are relatively easily recognizable – they have distinctive markings and ways of moving – and offer their services from a fixed location. Since the cleaners can be located again and again, the game is iterated. Hence, the mutual defection which characterizes one shot games is avoided. Instead, the fish play TFT.

The moral emotions

Contrary to what some discussions of evolutionary game theory suggest, TFT is no panacea. For one thing, if two players engage in it, it is vulnerable to mistakes: one defects by accident, and sets off a chain of mutual defections. ("Tit-for-tat killings" is, after all, how we often describe cycles of violence in the Middle East and elsewhere.) Moreover, it is a viable strategy only under the right conditions: nothing guarantees that the pay-offs of iterated interactions will be such as to make mutual cooperation in the interests of all players.

And even when the pay-offs are appropriate, TFT can break down. Psychologists who conducted experiments in which subjects played prisoner's dilemma against each other soon noticed that if the number of rounds was fixed in advance, cooperation tended to evaporate. The reason for this phenomenon is easy to see. Imagine a ten round game of prisoner's dilemma. We should expect the players to cooperate with one another in rounds one through nine, simply because they know that defection would set off a chain of mutual defections, and would lower each player's overall gain. But round ten is, effectively, a one shot game. There is no longer any point in cooperating, since the other player cannot retaliate on the next round. In a one shot game, as we know, defect is the dominant strategy; we can therefore expect each player to defect in the last round. But rational players soon learn that the last round is effectively removed from the iterated game. They therefore turn their attention to round nine, the next to last round. They quickly see that there is no point cooperating in this round: since they know the other player will defect in the next, they do not need to establish their goodwill. Hence, this round is effectively a one shot game as well. But with round nine removed from the iterated game, round eight becomes the last round, and the same reasoning applies. And so on. By making the game a fixed number of rounds, the incentive to cooperate is effectively removed.

Nevertheless, the conditions under which TFT is a viable strategy are sometimes met with in nature: interactions occur in which the pay-offs are appropriate, and the number of rounds is not fixed. In these circumstances, TFT turns out to be a powerful strategy, which can yield something like morality. Indeed, it may even explain the evolution of our sense of justice. In order for TFT to evolve, it must be possible for the players to be able, and be motivated, to detect cheats. As we have seen, vampire bats that refuse to feed other bats are refused food in turn, and therefore are less fit than those who cooperate. But it is not sufficient for bats to *behave* appropriately. This behavior does them no good unless it is recognized by others, that will one day be in a position to reciprocate. Thus bats will be motivated not only to be cooperative, but also to be *seen* to be cooperative. Indeed, this very fact provides them with an incentive to *appear* to be more cooperative than they really are. The better their reputation, the more likely it is that they will be fed, but if there is a cheaper way of raising their reputation than actually feeding other bats (which costs them a not insignificant six hours of time in which to find their next meal), they can be expected to take it. Thus, Trivers predicted, we should expect to find that sophisticated methods of cheating evolve: methods that do not involve the blatant refusal to cooperate but instead give the appearance of cooperating, without the substance. Once such mechanisms of cheating emerge, however, natural selection will place a premium on methods of detecting the cheats. We can expect to see an evolutionary "arms race," in which there is competition between ever more sophisticated methods of cheating, and ever better methods of detecting cheats.

One product of such an arms race might be our distinctive moral emotions. The moral emotions are comprised of two sets of partially overlapping feelings: the feelings we get in response to violations of moral norms, and the feelings which motivate us to live up to our moral obligations. For example, we feel certain types of anger in response to perceived injustices, while we feel guilt and shame in response to actual or projected wrongs on our own part. How can the evolutionary arms race between cheats and detectors give rise to these emotions? Trivers suggests that the selection pressures on both sides – both the pressures that favor more efficient means of cheating and those that favor more efficient methods of detecting cheats – might encourage the development of such feelings. One reason the moral emotions might be needed is to fill in the inevitable gaps in cheater detection systems. No matter how vigilant the members of a group are, opportunities for *free riding* will inevitably arise. A free rider, in the terminology of economists, is someone who enjoys the benefits which come from the provision of a public good – navigating her ship with reference to a lighthouse, driving on the roads paid for by taxes, or, least metaphorically, riding on public transport – without paying the cost. Free riding is rational, in a narrow sense of that word, when we have good reason to believe that we can get away with it. In the absence of the moral emotions, it might be rational more often.

Imagine a group of hominids in which there is a notion of moral obligation, but in which the moral emotions have not evolved. One member of the group free rides; perhaps he finds a food source which he ought to share with the others, but which, he believes, he can get away with saving for himself. He eats some, and hides the rest in a tree for later. All goes well at first, but on his last journey to the tree, he is followed by someone whose suspicion has been aroused by the free rider's unexplained absences. How should the members of the group react to this act of treachery? If they are – narrowly – rational, they will weigh the costs of punishing him against the benefits. On the one hand, allowing incidents of free riding to go unpunished sets a bad example, perhaps tending to increase the frequency with which it occurs. On the other hand, there may be significant costs to punishing the free rider. Sanctions will, presumably, have some element of coercion; they must involve, or at least be backed up by the threat of, physical force. But such force carries risks; even if the free rider is no stronger than average, in any physical confrontation both sides risk damage. And the free rider has just feasted; he is likely to be fitter and stronger than his would-be punishers! In addition, he might be a valuable member of the group, whose cooperation is needed in hunting or trading. Perhaps the appropriate sanction for his crime is banishment, but the group cannot afford to lose one of its members. All in all, though it might be rational to draw up lists of crimes and their appropriate punishments, when the time comes to punish those who transgress, it might be rational to do nothing.

If this is the case, then our little group has a problem. If it becomes known that free riding will not be punished, then the system of reciprocal altruism is in danger of breaking down. Without some kind of punishment, in effect, the group is no longer playing TFT but has reverted to "always cooperate." But we know that groups of cooperators are vulnerable to invasion by defectors. That is what we might expect to happen here: those members of the group who can avoid sanctions will free ride, whenever possible. As a result, they will be fitter than average, and their behavior will go to fixation. The dilemma facing our group is this: it is rational to promise to behave in certain ways, if certain conditions are realized, but it is not rational actually to behave in these ways in those conditions. It is rational to have systems of punishment in place to punish free riders, in order to discourage free riding, but very often when it comes right down to it, it is not rational to punish them! But if that is the case, then free riders will see through the threats of punishment, and behave as they like. A system of sanctions that will never be enforced is entirely superfluous.

The ordinary prisoner's dilemma presented us with a similar problem. There we saw that under the right conditions, being rational is not all that rational. If only the players in a one shot prisoner's dilemma were less rational, they might do better for themselves. Similarly, our group might be able to secure better outcomes for themselves if they were less rational. If they were so irrational as to carry out their threats of punishment *even if in doing so they hurt themselves*, then potential free riders would have a powerful incentive to refrain from transgressions. Emotions might plug this gap, allowing us to be less than fully rational on those occasions when it is rational not to be rational.[47] People – and perhaps some other animals – get angry, for example. Angry people are notorious for their inability to assess situations rationally; they often act to punish those who have made them angry without thinking of the costs to themselves. Imagine, now, our potential free rider, considering whether he should share his find with his tribe or hide it for his own consumption. He knows that his conspecifics are likely to get angry if they find him cheating, and that angry people do irrational things. The costs of cheating rise in this scenario, and our free rider might instead decide to share his find. Thus, the propensity to act irrationally in certain circumstances might be rational, in the sense that it is evolutionarily fitter.[48]

Other emotions might also be explicable in terms of their contribution to reciprocal altruism. Affection for certain group members, and antipathy to others, encourages us to play preferentially with those who reciprocate. This is especially the case if we know that these people will feel gratitude for our aid, which will motivate them to return it, and would feel guilt if they refused to reciprocate. Shame motivates wrongdoers to recompense those they have wronged, and thus tends to restore them to full status as members of the group, and therefore as potential partners in mutually beneficial exchanges. Sympathy motivates aid for others, and so on.[49]

This is good news for us: it seems to show that the emotions that underlie morality might be the product of natural selection. But the picture is not entirely rosy. If it is rational to possess the moral emotions, it is even more rational to *seem* sympathetic, and therefore trustworthy, than to actually *be* it. That way, you get the advantages which come from participation in exchanges, while remaining open to the possibility of defecting – without costly feelings of guilt or shame – if the opportunity presents itself. We might expect the ability to seem more "moral" than one really is to be strongly selected. Thus, we can expect animals who play TFT to be more strongly motivated to reciprocate when so doing is public, and therefore reputation-building, then when its pay-off is smaller. We can expect them to evolve methods of advertising that they are reciprocators, and that sometimes these adverts will be deceptive. But we can also expect that methods of detecting cheats will keep pace with these innovations. Trivers suggest that here we might have the origins of self-deception. There are characteristic cues which give us away when we are lying – sweaty palms, a quaver in the voice, and so on. These cues can be hidden, but as we all know from experience are hard to disguise. Far better, Trivers argues, if we are able to hide our own deceit from ourselves. If we believe that we are more moral, more prone to altruism than we really are, then we will be far more convincing in our attempts to deceive others. Self-deception might be an inevitable spin off of the profitable ability to deceive others.[50]

"Altruism" or altruism?

The concept of altruism with which we have been working is a technical one, drawn from the work of biologists. On this usage, a behavior counts as altruistic if it boosts the fitness of other organisms at a cost to the inclusive fitness of the agent. We have seen that biologists have powerful tools with which to demonstrate that much of what seems, in this technical sense, to be genuine altruism, is actually disguised genetic selfishness. When organisms aid their kin, they may boost their inclusive fitness. When they give help to other animals, even to members of other species, they may be engaged in reciprocal altruism; in helping others in expectation of a return that outweighs their costs. Even our moral emotions might be the product of genetic selfishness.

If that is all there is to morality, it seems that we shall be left with a bare simulacrum of it, shorn of its substantive content. Is morality no more than genetic selfishness? When people help one another, are they playing tit-for-tat: only assisting in the expectation of a return? This, we want to say, isn't morality at all. Strictly speaking, I do not act morally if I act only in my long-term interests, or even worse, in the interests of my genes. Altruism, genuine altruism, requires that we help others *for their sakes*, and not for our own.

We can easily convince ourselves that whatever else it is, reciprocal altruism is not genuine morality, by looking at the kind of downright *immoral* results to which it can give rise. If we act only in the interests of our genes, then we have little reason to aid those who will never be in a position to reciprocate. To be sure, we can think of some reasons to aid such people. Most obviously, aiding the indigent might be a good reputation-building strategy, if the aid is given publicly. By giving help when there is no possibility of reciprocation, we advertise our "altruism," thereby encouraging others to play TFT with us to our mutual benefit. Perhaps those who never give – or, more accurately, are never *seen* to give – to the indigent are regarded as potential defectors, and avoided by prospective partners in exchange.[51] Even secret giving can be explained using the resources biologists have available: if it is true that self-deception is an evolved adaptation, which disguises our selfishness from ourselves the better to hide it from others, then perhaps such apparently disinterested acts of charity are engaged in the better to convince ourselves of our own morality. In either case, the ultimate explanation is genetic selfishness. Altruism is revealed to be mere "altruism." The very fact that we need to disguise our true motivations even from ourselves shows how wide the gap is between our concept of morality, and the poor copy the biologists offer us.

Many philosophers and biologists have come to just such a melancholy conclusion. Once we realize that our moral emotions, supposedly our finest feelings, are the products of an evolutionary history in which long-term selfishness was systematically rewarded, we see that there really is no such thing as morality. Morality is – was *supposed* to be – about helping others for their sake, fairness and equity, equal consideration and justice. But it has its roots in selfishness and it bears the stamp of its origins.

In the light of these conclusions, some thinkers have gone so far as to argue that there really is no such thing as morality. It is a myth, as Richard Joyce puts it; and our belief that there is such a thing is unjustified.[52] He compares our situation to that of a paranoid person, John, who believes that Sally is persecuting him. Sally *might* be persecuting him, but given that we know that John is paranoid, we have little reason to rely upon John's testimony. He would believe that Sally was persecuting him, whether she was or not. Similarly, we would believe that there were moral facts and obligations whether there were or not: the dispositions to believe these things are built into our brains by evolution. Just as John's belief is unreliable because of his mental biases, so ours are unreliable because of ours.

Michael Ruse, a very prominent philosopher of biology, comes to similar conclusions.[53] Because the (alleged) existence of morality doesn't explain why we believe in morality, it is redundant. It plays no role in explaining our beliefs, nor our actions. Evolution explains both, not morality. This seems to me a mistake. The error lies in measuring morality against an inappropriate standard. It might be useful here to compare evolutionary explanations of morality with evolutionary explanations of belief in God. Several biologists have suggested that belief in God might be biologically adaptive.[54] Neuroscientists have even managed to locate the region of the brain that seems to play an important role in religious experience: the temporal lobe, which, when stimulated with powerful magnets, causes most people to have "God experiences."[55] Now, if this is true, it seems to me to be very bad news for theists. If belief in God is a product of evolution, and can be produced with the simplest brain manipulations, then we have little reason to place much faith in religious experience, whether our own or others. Given that most people would believe in God whether or not he existed, and would continue to have religious experiences in the absence of the divine, the actual existence of God seems redundant to an explanation of religious experience. As Michael Persinger, whose work on the temporal lobe opened up this line of enquiry, put it, neuroscience seems to show that "religion is a property of the brain, only the brain and has little to do with what's out there."[56]

This is exactly the claim made by Joyce and Ruse. Morality, they say, is a product of our minds, and has little to do with what is out there. But they miss a crucial difference between religion and morality. Religion is, precisely, concerned with what is "out there." If there is a God, his existence is entirely independent of our belief in him. If his existence is explanatorily inert, then this is very bad news for theists. But it is very plausible to think that morality is not like this. It is not independent of us and our beliefs, in the way in which God (and neutrons and giraffes and Italy) is. Instead, it is at least partially constituted of our beliefs and moral emotions. If pretty much all rational beings share a moral reaction (for example the strongly held belief that torture is wrong), and that reaction is a response to actual facts in the world (in this case, the suffering of victims of torture) then the fact that there is nothing *beyond* the feelings of observers and victims to refer to is neither here nor there. We have all we need to constitute moral facts.

The temptation to measure entities against inappropriate standards is a perennial one. It's the kind of temptation that led some philosophers to think that colors aren't *really* real. They realized that the colors we see are partially the product of our perceptual systems. For eyes like ours, (some) roses are red. For insects, they might be another color, or no color at all; under the sun of a different world, they might be brown or blue. So these philosophers concluded that colors weren't real: not *fully* real; as real as shapes and hardness and so on. In fact, the problem lay not with the colors, but with the tests they were required to pass before they counted as part of the furniture of the universe (as philosophers like to put it). If we all agree that roses are red – that they appear red to almost all of us, under conditions almost all of us agree count as normal – then they *are* red. We are mistaken if we think that colors are real in *the same way* as squares are, if we think that redness is simply "out there" like atoms and giraffes. But this is no reason to think that colors aren't real at all. So long as we can all agree upon them, and we have much the same experience of them, there is no reason to think that their ontological status is somehow lesser than those objects that exist regardless of our responses to them. The fact that we can use colors for such important tasks as controlling traffic demonstrates that we have no qualms about their existence.

Colors are a kind of thing that owe their existence to our perceptual equipment, as well as the physical features of colored objects. This does not make them illusory. We are not making a mistake when we say that roses are red. It is nothing like saying that moon is made of green cheese, or mistaking, from a distance, a gnarled tree for a person. Nor is it a merely subjective assertion (it is not like the statement "vanilla ice cream is the nicest," the truth of which varies from person to person). If we can count as mistaken in applying a predicate like "red," as we clearly can (I will usually be mistaken if I say that grass is red), and if we have clear and generally agreed upon conditions for the application of our color terms, there seems to be little reason to think that there is anything "iffy" about them.

But if we can say this for colors, then it seems that we can say it for morality as well. We have (relatively) clear criteria for the application of moral predicates. There are clear examples of moral mistakes ("torturing babies for fun is good") and generally agreed upon moral paradigms (it is not just Christians who recognize that Mother Theresa or Saint Francis of Assisi were good; not just Buddhists who recognize the goodness in Siddhartha Gautama). So what if our evolved capacities, dispositions and emotions play an ineliminable role in constituting moral goodness and badness? Our perceptual system, which is just as certainly the product of evolution, plays an equally significant role in constituting color. Since almost all of us – all except those who we rightly regard as suffering from an abnormality, whether it be psychopathy or blindness – share the same reactions, and since we have agreed-upon conditions for the application of our color (moral) terms, we have no reason to regard them as illusions.[57]

Moreover, just like our color perception, our moral reactions track real properties of events and people. Color perception reliably tracks the surface reflectant properties (roughly, the wavelengths of light reflected by the surface) of objects, so that objects appear different colors if their surface reflectant properties alter. It is harder to say what our moral reactions track, but it is clear enough that, at least to some extent, they track physical features of the world. They certainly are keyed quite closely into perception of suffering in others. The extent to which this is so – the extent to which we could describe all moral properties and events in purely physical terms – is a contentious issue among philosophers. But all sides agree on this much: the moral reflects physical features sufficiently so that the moral is *supervenient* on the physical, which is to say, whether or not we can adequately capture moral properties in physical terms alone, there are no moral differences between two situations unless there are physical differences. To this extent, we can be sure that our moral emotions are reliable guides to physical features of the world.

The philosophers who think that the evolutionary history of morality somehow undermines our usual conception of it are not yet done, however. They might claim that even if my argument shows that morality is not exactly an illusion, nevertheless, the picture of morality that emerges from it is very different from the picture I sketched at the beginning of this chapter, where we found that morality was a system of prescriptions, essentially concerned with the welfare of others, which were objective, unconditionally binding, and intrinsically motivating. Morality has a Kantian side, concerned with true beliefs, and a Humean side, concerned with motivations to action. But the defense of morality just outlined vindicates only part of the analysis. If morality is real in so far as, and because, the emotions that underlie it are real and generally shared, then only its Humean side is vindicated. Hume argued that morality was entirely to be explained in *subjectivist* terms; in terms, that is, of the emotions aroused in us by the contemplation of actions and states of affairs that we judge to be morally good or bad. We, like most evolutionary ethicists, have followed Hume in focusing upon the feelings that motivate actions, both ours and those of other animals, and in asking whether they are likely to include dispositions to care for the welfare of others. We have been concerned with morality as a *subjective* phenomenon, manifested in the feelings to which evolution might give rise. But this is not the whole of morality, as we have analyzed it. Morality, we said, was as much about belief as feeling. If we are forced, in the light of evolution, to give up on the cognitive side of morality, then the picture of the moral that will emerge will be radically altered. Evolution will have undermined not morality *per se*, but at least our commonsense concept of it.

Some philosophers have argued that this is just the route we should take. Michael Ruse is a case in point. Like us, Ruse locates the basis of the moral emotions in reciprocal and kin altruism (indeed, I have mined his work in developing the account of the origins of morality presented here). But Ruse is, explicitly, a Humean; for him, therefore, there is nothing more to morality than the kinds of emotions and dispositions which natural selection has implanted within us. He grants, as he must, that we cannot help but think that morality is objective, that it somehow transcends mere feelings. But the apparent objectivity of morality is, he claims, an illusion, foisted upon us by the same evolutionary process that gave us the moral emotions. We shall be more strongly motivated to act upon our desires if we believe that they reflect something beyond them, so the illusion of objectivity is functional, and has an evolutionary source. It remains, nevertheless, an illusion.[58]

But if we are forced to revise our concept of morality in the light of evolution, rejecting its cognitive side and the illusion of objectivity, we shall be left with a problem and a mystery. The problem is that, just as evolutionary theory might predict, people's moral emotions are much stronger and much more reliably triggered by close kin and members of one's community than by the more distant. People are often much more upset by a small slight to their parents or siblings than by a great injustice a thousand miles away. This is a problem, because many of us think that everyone ought to be given equal consideration, no matter where they are. In some of our moods, almost all of us think that the significant interests of the distant needy should outweigh the trivial interests of ourselves, our community and our close kin. Yet we continue to behave as if we didn't believe this: we spend money on extravagant presents for ourselves and for our family, and ignore (or donate little to) famines in Africa. If morality is just a matter of shared feeling, then perhaps it extends no further than the range of our reliably triggered moral emotions. Perhaps we are just wrong in thinking we have significant obligations to the distant needy. So Ruse, for one, concludes.[59] This is a problem, because it seems to involve the sacrifice of a very significant part of the content of our morality. If it is possible to retain this part, so that we can criticize others (and ourselves) when we neglect the distant needy and hope thereby to enlarge the scope of moral concern, this would be greatly preferable to following Ruse down the road of restricting the range of morality.

The subjectivist picture of morality, at least as developed by Ruse, has costs, both moral and conceptual. It also leaves us confronting a mystery: why is it that some of us almost all of the time, and most of us at least some of the time, have succeeded in expanding the scope of our moral concern beyond the targets which evolution predicts? Why is it that many people feel guilty when we remind them how many Ethiopian lives could be saved by the money they spend on chocolate bars or new shoes? Why is it that the circle of moral concern has grown over the past two hundred years, so that many people formerly excluded from it, or given little moral weight, have been included as full members of the moral community: people of all races, homosexuals, women, increasingly even animals? These changes have been too swift and too widespread to reflect genetic mutations. Instead, they are much more plausibly seen to be the upshot of moral *argument.* Sentiment *follows* conviction; it does not always lead it. It seems that there must therefore be a role for the cognitive elements of morality.

We can only make sense of moral argumentation, especially, but not only, as it is involved in the process of our expanding the sphere of moral concern, if we suppose that we engage in debate using our concept of morality as a constant reference point, *and* that our moral emotions are flexible enough to be shaped by the outcomes of our debates. We began to include blacks as full members of the sphere of moral concern, for instance, when rational arguments showed that there were no morally relevant differences between them and members of other races already in the sphere. It may well be that many people were intellectually convinced of this long before they responded appropriately, because intellectual conviction did not automatically engender emotional response. Eventually, however, most of us came to care (almost) as much about injustices to members of other races as to members of our own. The concept of morality had forced a revision in our moral responses.

But how did we come to have this concept? How did we go from having certain emotional responses to a range of acts and threats, to having a concept that we could then turn upon the very emotions which (presumably) engendered it? Here I get speculative. I suggest that the concept of morality is itself the product of evolution, and that we have come by it from an unexpected source: as an inevitable by-product of the development of that *im*moral phenomenon, self-deception.

Trivers's argument, you will recall, was that self-deception evolved because it was in our genetic interest to be taken in by our own claims of morality. Most of the time, it is in our interests to behave in accordance with the demands of morality upon us, but it is also in our interests to keep an eye out for occasions on which defection is profitable and we can get away with it without damage to our reputations. We want to *seem* moral without always *being* moral. But we'll be able to evade the increasingly sophisticated cheater detection mechanisms of our conspecifics only if we can fool *ourselves* into believing that we really are altruistic, that we sometimes act *for the sake of others*, and not merely because we expect a return. It seems to me that this hypothesis has an interesting implication. If we are to deceive ourselves, if we are to believe our protestations of selflessness, we must necessarily believe that morality is possible. Our acceptance of our own, perhaps false, claim to be a moral being requires that we possess the concept of morality. The notion we need here cannot be of morality as merely an adaptation, in any narrow sense, of a morality founded on reciprocal altruism, because our aim is precisely to convince potential reciprocators that we do not limit our concern to those who can benefit us. Trivers-style self-deception requires us to possess the *full* notion of morality, not its ersatz copy. It is altruism we must believe in, not "altruism." Thus, the idea of morality, the idea we have appealed to to criticize the copy foisted upon us by biology, might itself be the product of natural selection. We are evolved to believe in morality, the better to promote our own, narrower, concerns.

Once we have the concept, however, we are able to use it. We are able to judge our own and others' behavior against its standards, and not merely against those of genetic selfishness. We are able to begin the process of elucidation of the content of morality, the exploration of what it requires of us, in the Socratic manner: by examining our concepts. In this sense, moral philosophy might rest on an evolutionary basis. It is evolution that gives us our concept of morality, the very concept that we might utilize to criticize the genetic selfishness of evolution.

Stephen Jay Gould introduced the term *exaptation* in evolutionary biology. An exaptation is a characteristic of an organism which has been selected for because it fulfils one function, but which is then utilized for a quite different purpose. For example it has been suggested that feathers initially evolved due to their qualities as effective insulation: they enabled their possessors to regulate their body temperature more effectively. However, the animals that possessed them later found they could be put to different uses: they enabled gliding, and eventually flight. I am suggesting that morality might be an exaptation. We evolved a set of moral emotions, and, as a consequence, a conception of morality as objective and unconditionally binding. We then *exapted* this concept: using it is an independent measure of behavior. We turned it back against its origins. Morality, I suggest, might have had its source in the very self-deception we now condemn in its name.

Morality on other planets

I shall briefly consider one final evolutionary argument against moral objectivity. This one is inspired by reflecting on the fact that we humans may very well not be the only moral beings in the universe. Morality, full-blown, may have emerged on other planets. But what would alien morality look like? Some evolutionary ethicists argue that the kinds of actions that we regard as obligatory might be held to be immoral by some aliens. If their genetic constitution were different to ours or if their evolution took a different path, then the illusion of objectivity under which they labor might attach to actions we regard as immoral. Surely this is sufficient to show that objective morality is an illusion? If there were an objective morality, then it would be binding upon all rational creatures (as Kant pointed out). But there is no such morality.

What are we to make of this argument? The contention that the contents of our morality is sensitive to the details of evolutionary history is plausible. What counts as harming and benefitting someone, most obviously, is in important part a function of their biology, which makes them vulnerable to certain dangers and in need of certain resources and opportunities. But this fact is surely not sufficient to establish the species-relativism of morality. The fact that Australians are required to drive on the left hand side of the road, while Americans are required to drive on the right, does not establish any kind of interesting moral relativism. Similarly, the fact – if it is a fact – that, if evolution had taken a different path, we might have been required to eat one another's feces (to use Michael Ruse's rather off-putting example) is not sufficient to establish the species relativism of morality. It still might be the case that, considered at a high enough level of abstraction, all beings would evolve the same morality. It would differ in its specific injunctions ("eat up all your brother's feces!") but its most general principles would be just the same ("treat others as you would have them treat you").

What if evolution had taken a *radically* different path, so that not only might different kinds of actions harm and benefit different kinds of creatures, but there might even be different *kinds* of creatures? Would this be sufficient to establish a species-relativism strong enough to refute moral objectivity? At least one philosopher has argued that it would (and that reflection on this fact should be sufficient to undermine the illusion of objectivity, even in the absence of evidence that evolution actually has unfolded differently elsewhere in the universe). Waller asks us to consider a kind of creature something like the Borg in *Star Trek*: as intelligent as *Homo sapiens*, but with no notion of individuality, perhaps because of a haplo-diploid chromosomal arrangement like ants and bees. These creatures would have fundamental moral obligations quite different from ours, and would find our emphasis on the individual and her rights "not merely absurd, but morally odious".[60]

I am not sure that the discovery that our morality is merely local should cause us to question its objectivity; it might be that local objectivity is objectivity enough. Be that as it may, I am in any case unconvinced that Waller's thought experiment succeeds in showing that our morality is local. Who is the subject of moral obligations, in Waller's thought experiment? He seems implicitly to assume that each ant-like entity is comparable to each one of us. But this seems to me a mistake. It might be better to think of the entire community of ant-like entities as an individual being, so that the community would be the appropriate subject of rights and obligations. If this is correct, then the fact that each ant-like entity has no rights against the community is no more interesting than the fact that my skin cells have no rights against me: they are part of me, and are appropriately sacrificed for the greater good. If, on the other hand, each ant-like entity should be conceived of as an individual in its own right, then we can insist they are making a mistake in rejecting our notion of individual rights.

Evolved morality is real morality

Ruse, Waller, and the other evolutionary deflationists are half right. They are right in thinking that morality must have its origins in (genetic) self-interest; it could not have been selected for otherwise. They are also right in thinking that evolution alone could give us the Humean side of morality, its subjective and motivational side. However, they are wrong in thinking that unless we reject philosophical naturalism altogether, and accept a supernatural source for morality, we need to conclude that that is all there is to morality. Instead, evolution is capable of endowing us with the notion of morality as an objective system, and providing us with the means of acting in accordance with it. That is to say, evolution can account for the origins of a morality which meets all the conditions of our analysis: both its Humean, motivational, side, and also its broadly Kantian, objective side.

Morality comes to us as a product of our evolutionary history. This history systematically favored (genetic) selfish behavior and eliminated genetic altruism. Yet it gives us the very concept that leads us to condemn selfishness and approve of selflessness. Evolution provided us with a concept we can turn back against evolution. From the mindless and mindlessly selfish rose beings capable of rationality and morality.

Throughout this book, I shall be concerned with steering a middle course between those thinkers who deny the significance of evolution (and more generally our biology), for thought and morality, and those who claim that we can capture everything that is significant about human beings in essentially biological terms. As we can already begin to see, both sides capture part, but only part, of the truth. We have morality only as a consequence of our evolutionary past. Moreover, our morality continues to bear the clear traces of that past in it, and it is reasonable to think that it always will, no matter how long we human beings survive. However, the morality we have, today, is very different from the core of proto-morality we share with vampire bats and cleaner fish. It is more extensive and more demanding, as a result of millennia of rational elaboration of its content. Children today inherit this morality from their parents and from their culture. They are *taught* an ethic of equal concern. Our inheritance from past generations is not only via our genes, but also through our socialization and our education. As a result, the intellectual and moral development of each child follows a quite different route, to a quite different destination, from that of its hunter-gatherer ancestors. Our evolutionary past constrains what we can think and believe and hope for; equally, it opens us up to unexpected, and ever-changing, vistas of transformation and (we can hope) progress.