# 1 The Natural Selection of Helping

## 1.1 Altruism and Selfishness

What is the value of moral thinking? What benefits do we get from engaging in this way of judging ourselves and each other? It is reasonable to assume that the answer has something to do with helpfulness and cooperation. In some vague manner we expect that a person who thinks of human interactions in terms of "virtue," "obligation," or "justice" is more likely to be a useful member of society than someone for whom these concepts are entirely alien. This natural thought will be examined in detail in chapter 4; I mention it now because one might suppose that it represents an immediate barrier to the hypothesis that morality is innate. Surely, one might think, natural selection is a competitive race where the laurels go to the untamed individualist? The main goal of this first chapter is to show that this doubt is misguided. I will outline several means by which helpful, cooperative traits may evolve. But before embarking on that task we should get our thinking straight on a misleading piece of terminology. Much of the literature on the topic of how helpful behaviors might naturally emerge professes to concern the evolution of *altruism*, but this is a word that has, over the years, been muddied and fudged in discussions about evolution, leading to persistent confusion and erroneous conclusions in some quarters. In the context of this book it is important to be exact about what we mean in this respect, and so I will begin with a three-way distinction.

*Helping* Behaving in a way that benefits another individual. Contrast: harmful behavior. (I am happy also to call this "cooperation" or "prosocial behavior.")

*Fitness sacrificing* Behaving in a way that advances another individual's reproductive fitness, at the expense of one's own reproductive fitness. (This is often called "evolutionary altruism.") Contrast: fitness advancing.

*Altruism* Acting with the intention of benefiting another individual, where this is motivated by a non-instrumental concern for his or her welfare. Contrast: selfishness.

In restricting "altruism" in this way I take myself to be respecting ordinary English. In English, an action is altruistic only if done with a certain other-oriented deliberative *motivation*, in the sense that the welfare of another was the agent's ultimate reason for acting. Suppose Amy acts in a way that benefits Bert, but what motivates the action is Amy's belief that she will benefit herself in the long run. Then it is not an altruistic act but a selfish act. Suppose Amy's belief turns out to be false, so that she never receives the payoff and the only person who gains from her action is Bert. Does this cause us to retract the judgment that Amy's action was selfish? No. Whether an action is selfish or altruistic depends on the deliberative motivating reasons for which it was done—the considerations in light of which it was performed[1]—not on who ends up benefiting from its performance. Some people doubt whether *any* human actions are altruistic in this respect; they think that all actions are done from the ultimate motive of self-gain. As we will see later, these people (psychological egoists) are almost certainly wrong. Many human actions are done from a genuine regard for others, and are not motivated by considerations of self-gain.

There are few non-human animals, if any, that can be spoken of uncontentiously as having motivating reasons—in the sense of having considerations that figure in their deliberations—and certainly there are few non-human animals that have the concept of *self* necessary to have a selfish motivation. Therefore there are few non-human animals, if any, that can be spoken of uncontentiously as being altruistic or selfish. I am not going to define with any precision what a motive or intention is, nor am I going to attempt to gauge where to draw the line between creatures that have them and those that don't. It is enough to point out that, say, plants do not have the requisite sort of motives and intentions, and so, although plants may behave in ways helpful to others, and perhaps they may behave in fitness-sacrificing ways, I judge it best to eschew describing a plant's behavior as "altruistic" or "selfish."

Quite obviously organisms help each other, though there are some difficulties that would have to be straightened out before we could be satisfied that we had nailed down the notion fully.[2] Presumably "benefiting X" can be treated as synonymous with "advancing X's interests," but worries arise concerning what it takes to "have an interest." (Even when we are dealing with humans, who clearly have interests if anything does, things are far

from straightforward.) If one squashes a cockroach on the kitchen floor, there would be nothing unusual in claiming that doing so "frustrated its interests" (and in saying this we need not be indulging in any dubious anthropomorphism according to which the cockroach has conscious desires, motives, or experiences pain), but it proves difficult to say what a cockroach's interests amount to. Perhaps in the end the notion cannot be explicated, and such appeals are confused. What is important is that we don't unthinkingly equate a creature's interests with its reproductive fitness. A creature's interests *might* coincide with its reproductive fitness—this might be how we end up explicating cockroach interests—but it certainly doesn't need to. When I buy a birthday present for my friend, this is a kind of helping behavior (even if it is only pleasure I'm trying to bring). But it is *her* interests that I am seeking to advance—the individual who is born, lives, and dies—not those of her genes. Talking of "the interests of genes" is even more shaky than talking about the interests of a cockroach. Genes don't really have an interest in replicating, any more than a river has an interest in not being dammed. But that a gene has the characteristics that it does have is explained by the contribution those characteristics have made to its successful replication and endurance; thus we can speak—in a quasi-metaphorical way—of a gene being "designed" to replicate, of replication being its "purpose." And if we allow ourselves to talk this way, then speaking of a gene having "an interest" in replicating is hard to resist.

If we give in to this temptation, and allow ourselves to talk of my friend's genes having interests as well as of *her* having interests (or, if you prefer, of her "genetic interests" versus her "individual interests"), it is clear with whose interests I, as a friend, am concerned when I choose a birthday present. It may be a year's supply of contraceptives that I give her for her birthday, thus preventing the replication of her genes, but it is no less helpful for that. Conversely, it is *possible* that I might be concerned with the interests of her genes (though that would be pretty strange), in which case I might perform the action of secretly sabotaging her supply of contraceptives. But in advancing my friend's genetic interests in this way I would hardly be acting helpfully to *her*. Similarly, an act of patricide or matricide may advance the genetic interests of the perpetrator, and thus also those of the murdered parent, but it is no less harmful to the victim for that! And there is nothing incoherent in the idea of a person's interests being ruined by her being forced to have a large family. In short, to confuse a person with her genes is as silly as confusing her with her lungs or her lymph nodes, but as soon as the distinction is enforced, so too must be the distinction between a person's interests and the "interests" of her genes.

I admitted that when we get down to organisms that cannot experience anxiety or pain it *may* be permissible to identify the organism's interests with the advancement of its reproductive fitness. A useful thought experiment is to ask ourselves whether our view about harming an individual organism would change if we were to find out that it is sterile and not in a position to give aid to any kin in their reproduction. In such circumstances we can do no harm to the organism's reproductive fitness; thus, if we still feel comfortable saying that squashing or killing the organism "harms it" (as, I submit, we usually do), we must be employing a different notion of "harm" or "interest": one that pertains to the harming of *the individual*.[3] (Likewise with helping the organism.)

There can be little doubt that "fitness-sacrificing" behavior occurs. People sometimes give up on their plans of raising a family in order to devote themselves to charity work in distant countries. There are cases of people in times of war committing suicide in order to save the lives of their comrades. If their comrades were family members, then such acts of heroic sacrifice might still count as "fitness-advancing" behavior; but in many cases the beneficiaries of the sacrifice are unrelated. What *is* controversial about fitness-sacrificing behavior is not whether it occurs, but whether it might be selected for by the forces of natural selection. Some people argue that fitness-sacrificing behavior cannot be favored by natural selection. Richard Alexander (1987: 3), for example, asserts that there is "not a shred of evidence" that such behavior is a "normal part of the functioning of every human individual." A corollary of this might be the claim that non-accidental fitness-sacrificing behavior cannot be found outside the human species.

We will see later in this chapter that, on the contrary, fitness-sacrificing behavior might well be produced by biological natural selection. The key is that the population upon which natural selection works is structured in a certain grouped manner. But in fact this is not a dispute that matters here; relative to the aims of this book, what matters is that certain kinds of *helpful* behavior have been selected for in humans. Whether these helpful behaviors are also fitness sacrificing, or whether they are really a form of fitness advancement, is something I am content to leave open. The question in which I am interested is "What proximate mechanisms might be favored by natural selection in order to regulate this helpful behavior?" One possible answer, which I think is correct, is "Altruism." In other words, in order to make an organism successfully helpful, natural selection may favor the trait of acting from altruistic motives (assuming the organism has the cognitive sophistication to have motives at all). Another possible answer, which I also think is correct, is "Morality." In order to make an organism

successfully helpful, natural selection may favor the trait of making moral judgments. Exploring this second answer is the main task of this book.

The goal of the rest of this chapter is to identify the principal evolutionary processes that may lead to helpful organisms. This will put us in a good position then to ask whether a moral sense may have developed in humans as a means of governing helpfulness. But first a further cautionary word. Suppose that the above approach leads to a positive outcome, and we decide that human moral thinking is governed by dedicated mechanisms that evolved through the process of Darwinian selection. The conclusion that would be absolutely incorrect to draw is that what these arguments show is that all human action, even what is helpful and what is deemed morally virtuous, is "really selfish." Drawing the distinctions above should be sufficient to show what is wrong with this conclusion; but the tendency to leap to this assumption appears to be so persistent, and is so pernicious, that it pays to underline the error. Richard Dawkins (1981: 3) concludes, on the basis of his "selfish gene" view, that "we are born selfish." Alexander (1987: 3) writes that we will not understand human conduct until we grasp that societies are "collections of individuals seeking their own self-interest." And Michael Ghiselin (1974: 247) memorably tells us "Scratch an altruist, and watch a hypocrite bleed." But such attitudes, posing as hard-nosed realism, erroneously conflate distinct explanatory levels. (See Tinbergen 1963.) In particular, they commit the basic blunder of confusing the cause of a mental state with its content. If a person's nervousness about a pending job interview is partially caused by the fact that he just drank four cups of strong coffee (had he not drunk the coffee, he wouldn't now be nervous), it would be crazy to conclude that really he is nervous *of the coffee*! Yet people who think that evolutionary explanations reveal the "true" content of all our motivations, reasons, and interests fall foul of exactly this piece of mistaken reasoning. Suppose Fred is looking after his sick wife. When asked why he does so, he reports sincerely that he wishes to alleviate her suffering for her sake, because he loves her. An evolutionary psychologist might then tell us that it is to Fred's reproductive advantage to look after his spouse, for then he will have help in raising his offspring, adding that the love that Fred feels for his wife is the output of a proximate mechanism by which natural selection ensures that a person helps his mate when she needs it. Thus, an evolutionary explanation has been provided for a cognitive/emotional/behavioral phenomenon: Fred's love for his wife. But this explanation reveals nothing about the content of Fred's motivations, and doesn't show that he "really" cares about his reproductive fitness and only derivatively cares about his wife's welfare.

One might object that there is a disanalogy here. In the case of an evolutionary psychological explanation, it might be thought, the explanans that is appealed to is something that (unlike a cup of coffee) itself has interests: genes. But we have already seen that talk of genes literally having interests is shaky stuff. (I think the metaphor causes more confusion than it's worth—exactly the kind of confusion an author has to waste time combating in the first chapter of a book on the evolution of morality.) But even if we were to earn the right to that kind of talk at a literal level, the argument would still be unsound. It would require the endorsement of the following "principle of interest transferal":

If X has interests *a*, *b*, *c*, etc., and X having those interests is explained by the fact that Y has interests *p*, *q*, *r*, etc., then X's interests are "subservient" to Y's, and in fact X's "real" or "ultimate" interests are *p*, *q*, *r*, etc.

But there is no reason to believe in this principle, and good reason to reject it. It continues to confuse *explaining the origin* of a mental state (or interests) with *providing the content* of that state (or interests). The source of this common confusion may be an ambiguity in the notion of "a reason." *Fred's* reason why he cares for his wife is her suffering. This is what motivates him and figures in his deliberations. *The* reason why her suffering motivates him (or, better, *a* reason) may be that caring for one's partner advances one's fitness, and thus has been selected for in humans, and Fred is a human. When we explain a person's behavior and mental states by appealing to the fact that his genes have replication-advancing characteristics, we are giving reasons for his having these mental states and behaving in this way. But to conclude that these are therefore *his* reasons—the considerations in light of which he acts—is a gross mistake. In exactly the same way, we can wonder about the reason that an avalanche occurred, but in doing so we are hardly wondering about what malicious motives the melting snow harbored.[4] I am not claiming that a person's reasons must always be obvious and apparent to her; all I am saying is that they are not all "ultimately" concerned with genetic replication.

The three categories of action that have been identified in this section can be conjoined in any combination. Ignoring for a while those cynics who deny the existence of altruistic behavior, we can come up with examples satisfying all of the following conditions[5]:

- behavior that is helpful, fitness sacrificing, and altruistic
- behavior that is helpful and fitness sacrificing, but selfish
- behavior that is helpful, fitness advancing, and altruistic
- behavior that is helpful, but fitness advancing and selfish

- behavior that is unhelpful, but fitness sacrificing and altruistic
- behavior that is unhelpful, fitness sacrificing, and selfish
- behavior that is unhelpful and fitness advancing, but altruistic
- behavior that is unhelpful, fitness advancing, and selfish.

With these distinctions made and some potential confusions nipped in the bud, we can turn to the first step of the argument. Note that our focus is not on altruism—either in the vernacular psychological sense or in the evolutionary sense (which I have been calling "fitness-sacrificing behavior"). Altruism is not an important issue in this book. Nor is our focus, in the first instance, on "moral" behavior. Rather, our initial task—the task of the rest of this chapter—is to outline the evolutionary processes that may lead to the development of *helping behavior*. In due course we will ask the question of whether moral thinking may be a mechanism that in humans regulates helping behavior—but noting this is to look downstream. Issues pertaining to morality are what we are working toward, but they will not surface in our discussion for a while.

Just as there are many evolutionary reasons for organisms having the capacity for locomotion, say, there may be many evolutionary reasons for organisms having the capacity to help each other. My objective, then, is not to alight upon *the* way in which helping behavior is selected for, but to sketch some of the broad evolutionary forces: kin selection, mutualism, reciprocity, and group selection. Lastly I will discuss how culture may have affected helping traits in the special case of humans.

## 1.2 The Evolution of Helping: Kin Selection

Always the first to be mentioned is what is usually called *kin selection*, the locus classicus of which is William Hamilton's 1964 paper "The Genetical Evolution of Social Behaviour" (though it was a selective force vaguely appreciated by Darwin[6]). It helps if we think, as Richard Dawkins has famously urged us to, of organisms as vehicles by which genes succeed in reproducing themselves. An organism that is kind and helpful to its family members—that is, to those that are guaranteed to share its genes—may be a useful sort of vehicle for a gene to inhabit. As far as the gene is concerned, if its "vehicle" sacrifices its life to save three offspring, or three siblings, or nine cousins, then that's a good deal. Talking of life sacrifices is a bit dramatic; we're just as much concerned with more modest sacrifices: sharing food with your siblings, looking after your young nieces and nephews, educating your own children. That a creature should care for its own offspring

is so engrained in our minds that it takes some effort to attain the critical distance needed to realize that it requires some explanation in evolutionary terms. Many creatures don't care for their offspring, preferring to opt for quantity over quality. But most mammals go for quality offspring, and this requires the provision of a degree of caring, feeding, and nurturing. A human infant is remarkably dependent on the help of others, and remains so for many years. Therefore we should expect that the trait of caring for one's children has been strongly selected for in humans. A gene inhabiting a human vehicle that wasn't inclined to care for its children—that left them to fend for themselves upon birth—would quickly become history.

Consider the Hymenoptera class of social insects: ants, bees, and wasps. We can point out three interesting features of such insects. First, they are paragons of social success. If any group of individuals can be said to count as a "super-organism," it is an ants' nest or a beehive. Second, they manifest an unusual amount of helpful behavior. Bees have their suicidal sting, which they use in defense of the hive. There are castes of ants in the nest that are born sterile, and spend their days tending the offspring of others. Evolutionary theory needs to explain these peculiarities. How could the trait of sterility or a suicidal tendency possibly evolve by biological natural selection? Wouldn't natural selection favor the bee that *doesn't* sting? Indeed, Darwin recognized that the social insects raised a problem "which at first appeared to me insuperable, and actually fatal to my whole theory" ((1859) 1998: 352). Hamilton was able to provide an answer unavailable to Darwin, by drawing attention to the third peculiarity of these insects: their genetic interrelatedness. In a nest of ants, bees, or wasps, many of the individuals are much more closely related in genetic terms than in a group of, say, monkeys, groundhogs, or humans. The mammals that we are familiar with share, at most, 50 percent of their genetic material with their immediate family members (identical twins notwithstanding).[7] But things are different with the Hymenoptera, due to their unusual chromosomal arrangement: The male bee has half the number of chromosomes as the female bee, and the female "sisters" of the nest (by far the majority of nest members) share 75 percent of their genetic material with each other.[8] Hamilton's Rule states that a trait of helping others at some cost to the individual can be expected to be favored by natural selection if

$$rB > C,$$

where $r$ is the degree of genetic relatedness to the individual, $B$ is the benefit to the recipient, and $C$ is the cost to the individual. In the Hymenoptera, $r$ is often higher than it is with mammalian conspecifics, allowing $C$ to be

proportionally higher. Given this unusual circumstance, we would predict a greater amount of sacrificial helping behavior in bees than in, say, mice—and this is precisely what we do observe.[9]

So we have a perfectly plausible and highly confirmed theory about why and how biological natural selection produces organisms who help out family members at some cost to themselves. Yet it seems that this could hardly explain human *morality*, in which (at least in the Western tradition) the tendency to favor one's own family members is a vice to which we have given the name "nepotism." Moreover, kin selection seems unable to explain the evolution of helping behavior toward non-kin, which, clearly, is an important element of human morality. If two creatures are unrelated, then $r$ in Hamilton's Rule will be zero, and thus so will be $rB$, and thus kin selection will be unable to explain any helpful behavior for which $C > 0$—that is, helpful behavior that is in any way costly to perform.

Yet kin selection may still be an important factor in explanations of helping offered to non-kin. First, it should be borne in mind that the trait of helping kin must involve proximate mechanisms that allow organisms to recognize kin, and these mechanisms may be sufficiently fallible—especially in novel environments—that they prompt helping behavior towards non-kin. In many species kin recognition is achieved via scent; for example, nurturing behaviors in the parent may be triggered by the odor of the newborn activating hormonal responses (Yamakazi et al. 2000). But kin-identifying mechanisms may be much coarser than this. If the population is structured in small family groups such that the conspecifics with which an individual most frequently interacts are very likely to be kin, then natural selection could plump for a simple solution: "Provide help to those conspecifics with whom you interact frequently." A good example of nature using such coarse-grained mechanisms is how hatchling chicks "imprint" on the first object they see moving, be it a human or a rotating red cube. (See Lorenz 1937; Bateson 1966.) In the natural environment the mechanism works well enough, since the first moving object seen is nearly always the mother. We can find evidence of this sort of phenomenon in humans too. Studying people raised on kibbutzim, the anthropologist Joseph Shepher (1971, 1983) found that there is a strong tendency not to be sexually attracted to any individual with whom one was raised, irrespective of whether he or she is genetically related. The hypothesis (which had been put forward by Edward Westermarck in the nineteenth century) is that this is a mechanism for incest avoidance. Natural selection does not make humans avoid sibling incest by developing a "sibling detector"; it prefers the simpler "familiar-from-childhood detector." In the ancestral environment,

the two mechanisms would pick out pretty much the same extension of individuals, and the latter has lower running costs. (For modern confirmation, see Lieberman et al. 2003.) If, then, human helping towards kin (or certain classes of kin) is regulated by a "provide-help-to-those-conspecifics-with-whom-you-interact-frequently" mechanism, and humans now live in societies in which we interact with far more conspecifics than natural selection ever dreamed of (including the "virtual interactions" supplied by TV, newspapers, and so forth), then one would expect to observe, ceteris paribus, a great deal of helping behavior towards non-kin, despite the fact that kin selection is the only explanatory process in play.

A second reason why kin selection may be important regarding helping behavior toward non-kin is that in it we at least have an explanation for how and why certain creatures will have in place the mechanisms that regulate helpful behavior. Biological natural selection is a conservative process, bending old structures into new, pressing into service available material for novel purposes. For example, the hormone that in mammals seems to govern maternal nurturing behavior is oxytocin—an ancient hormone, found even in mollusks, that was co-opted for the job more than 200 million years ago (Allman 2000: 97, 199). We now know that oxytocin is also centrally involved in pair-bonding behavior, suggesting that natural selection has tweaked its role over millions of years in order to encourage more extensive helpfulness beyond the mother-offspring bond. (See Gimpl and Fahrenholz 2001; Uvnäs-Moberg 2003.) If kin selection gave our distant ancestors the psychological and physiological structures needed for regulating helpful behavior toward family members, then those structures became available for use in new tasks—most obviously, helpful behavior toward individuals outside one's family—if the pressures of natural selection pushed in that direction. And there are several ways in which they may have pushed in that direction.

## 1.3 The Evolution of Helping: Mutualism

Sometimes there are ends that would benefit a creature but which it cannot achieve alone. A lion might want a piece of elephant for dinner—or there may be nothing else available—but one lion will not be able to accomplish this by itself. If a group of lions find themselves in this situation, they will do well by cooperating in the bringing down of an elephant. If they don't cooperate, all of them will go hungry; maybe if they don't cooperate, all of them will die. Even if it's not an elephant that is on the menu but something that a lone lion might stand a chance of capturing, by hunting

together the lions vastly improve the probability of success and lower the risks. Clearly, such lions do not need to be genetically related in order for natural selection to push in favor of traits that encourage this kind of cooperative behavior. This kind of helping is sometimes called *mutualism* and sometimes *cooperation*. However, in ordinary contexts the word "cooperation" can be applied to many other kinds of mutually beneficial arrangements as well (such as reciprocal exchanges, which are to be discussed next), and so employing it in a restricted sense is apt to lead to confusion. "Mutualism" is sufficiently unfamiliar outside its theoretical context that it is the preferable word.[10]

Helping behavior that is explained by reference to mutualism is not fitness-sacrificing behavior. A lioness who doesn't cooperate threatens to spoil the whole hunt (for herself as well as the other lions), and thus lowers her own reproductive fitness. There may be circumstances where the participation of *all* the lions is not needed to bring about the desirable end, and in those circumstances there will be a selective pressure upon lions to hang back and let others do the work. But given that, as a general rule, the more lions are involved in the hunt, the higher the probability of a successful kill (and the lower the probability of any hunter getting hurt), often joining in the hunt will be a better means of advancing reproductive fitness than not doing so.

One feature of mutualism to which it is important to draw attention—in order to contrast it with the next process of helping—is that it does not require ongoing relationships among the participants. For example, when a group of small birds "mob" a large threatening animal in order to drive it off (another example of mutualistic helping), they have each advanced their fitness right then and there, and this fact wouldn't alter if all the birds then dispersed and never interacted again. So mutualism is not a reciprocal relation. The simple difference can be brought out using a nice example from David Hume, who imagines two oarsmen pulling together in order to row to a destination each desires. No promises are exchanged between the two, for none are needed: If either stops rowing, the boat will go in circles; it is a situation in which the desired end will be reached only "if all perform their part, but loses all advantage if only one perform" ((1751) 1983: 95). We can imagine a comparative case involving the kind of rowboat that can be propelled by a single oarsman. If one person promises to take the other to her destination if she agrees to row him to his destination later in the day, then this is a different kind of arrangement, involving a kind of contract (though perhaps a tacit one). The former case is an example of mutualism, the latter of reciprocity. Let us now turn to reciprocity directly.

## 1.4 The Evolution of Helping: Direct Reciprocity

It is a simple fact that one is often in a position to help another such that the value of the help received exceeds the cost incurred by the helper. If a type of monkey is susceptible to infestation by some kind of external parasite, then it is worth a great deal to have those parasites removed—it may even be a matter of life or death—whereas it is the work of only half an hour for the groomer. Kin selection can be used to explain why a monkey might spend the afternoon grooming family members; it runs into trouble when it tries to explain why monkeys in their natural setting would bother grooming non-kin. In grooming non-kin, the benefit given by an individual monkey might give a great deal more benefit than cost incurred, but still the groomer incurs *some* cost: That half-hour could profitably be used foraging for food or arranging sexual intercourse. So what possible advantage could there be in sacrificing *anything* for unrelated conspecifics? The obvious answer is that if those unrelated individuals would then groom *her* when she has finished grooming them, or at some later date, then that would be an all-around useful arrangement. If all the monkeys entered into this cooperative venture, in total more benefit than costs would be distributed among them. The first person to see this process clearly was Robert Trivers (1971), who dubbed it *reciprocal altruism*.

One of Trivers's primary examples of these values working out in favor of helping exchanges is the "cleaning stations" on a coral reef. Small "cleaner fish" (or shrimp) indicate their willingness to remove from a large fish its external parasites by approaching the host with a distinctive swimming pattern. The large fish, if it wants cleaning, responds by opening its mouth and gill plates in order to allow the cleaners to go to work. When the host has had enough, it gives a distinctive signal to this effect, and the cleaners depart. The host fish could, on any given occasion, get a cleaning *and* an easy meal at the end of it. If the undersea world were teeming with willing cleaner fish, then perhaps it should do just that. But given that the reef will support only so many groups of cleaners, it is to the large fish's fitness advantage to keep this exchange going. It then knows where to go for a good cleaning, it knows that these are reliable cleaners, and that's worth something. It's worth more than a free meal. If the big fish gives up a free meal for long-term benefit, what do the cleaner fish give up? By approaching the large fish—entering its mouth even—they take a risk; so they give up safety. It's actually impossible for the cleaner fish to gain their benefit (eating ectoparasites) without paying this price. However, they could still "cheat" by taking small mouthfuls out of the unsuspecting large

fish's fins (as certain species of "cleaner mimics" do), thus increasing their immediate net gain at the other's expense. But this would be a myopic choice. Just as it is hard to find a good cleaner, so is it hard to find a loyal customer.[11]

A relationship whose cost-benefit structure is that of reciprocal altruism could exist between plants: organisms with no capacity to cheat, thus prompting no selective pressure in favor of a capacity to detect cheats. Even with creatures who have the cognitive plasticity to cheat on occasions, reciprocal relations need not be vulnerable to exploitation. If the cost of cheating is the forfeiture of a highly beneficial exchange relation, then any pressure in favor of cheating is easily outweighed by a competing pressure against cheating, and if this is reliably so for both partners in an ongoing program of exchange, then natural selection doesn't have to bother giving either interactant the temptation to cheat, or a heuristic for responding to cheats. But since reciprocal exchanges will develop only if the costs and benefits are balanced along several scales, and since values are rarely stable in the real world, there is often the possibility that a reciprocal relation will collapse if environmental factors shift. If one partner, A, indicates that he will help others no matter what, then it may no longer be to B's advantage to help A back. If the value of cheating were to rise (say, if B could possibly *eat* A, and there's suddenly a serious food shortage), then it may no longer be to B's advantage to help A back. If the cost of seeking out new partners who would offer help (albeit only until they also are cheated) were negligible, then it may no longer be to B's advantage to help A back. For natural selection to favor the development of an ongoing exchange relation, these values must remain stable and symmetrical for both interactants.[12] What is interesting about many reciprocal arrangements is that there is a genuine possibility that one partner can cheat on the deal (once she has received her benefit) and get away with it. Therefore there will often be a selective pressure in favor of developing a capacity for distinguishing between cheating that leads to long-term forfeiture and cheating that promises to pay off. This in turn creates a new pressure for a sensitivity to cheats and a capacity to respond to them. An exchange between creatures bearing such capacities is a *calculated* reciprocal relationship; the individual interactants have the capacity to tailor their responses to perceived shifts in the cost-benefit structure of the exchange (de Waal and Luttrell 1988).

The cost-benefit structure of a reciprocal relation can be stabilized if the price of non-reciprocation is increased beyond the loss of an ongoing exchange relationship. One possibility would be if individuals actively punished anyone they have helped but who has not offered help in return.

Another way would be to punish (or refuse to help[13]) any individual in whom you have observed a "non-reciprocating" trait, even if you haven't personally been exploited. One might go even further, punishing anyone who refuses to punish such non-helpers. The development of such punishing traits may be hindered by the possibility of "higher-order defection," since the individual who reciprocates but doesn't take the trouble to punish non-reciprocators will apparently have a higher fitness than reciprocators who also administer the punishments. Robert Boyd and Peter Richerson (1992) have shown that this is not a problem so long as the group is small enough that the negative consequences of letting non-reciprocators go unpunished will be sufficiently felt by all group members. They argue, however, that we must appeal to cultural group selection in order to explain punishing traits in larger groups. (More on this in section 1.7.)

Two important things need to be noted. First, these "reciprocal altruists" are not *altruists* in the sense that I have defined it. The example, after all, is of types of fish, which do not satisfy the psychological prerequisites for performing actions that are either altruistic or selfish in the vernacular sense of these words; they may not satisfy the prerequisites for performing *actions* at all. Second, and perhaps less obvious, these helping organisms are not exhibiting fitness-sacrificing behavior either (therefore they are not "evolutionarily altruistic"—see Sober 1988). In a reciprocal exchange neither party forfeits fitness for the sake of another. As Trivers defined it, "altruistic behavior" (by which he means *helpful* behavior) is that which is "apparently detrimental to the organism performing the behavior" (1971: 35)—but obviously an *apparent* fitness sacrifice is not an actual fitness sacrifice, any more than an apparent Rolex is an actual Rolex. Others have defined "reciprocal altruism" as fitness sacrificing *in the short term*. But again, forgoing a short-term value in the expectation of greater long-term gains is no more an instance of a genuine fitness sacrifice than is, say, a monkey taking the effort to climb a tree in the hope of finding fruit at the top. So despite claims that reciprocal altruism and kin selection together solve the so-called paradox of evolutionary altruism, if (i) by "altruism" we mean *fitness sacrificing* (not *apparent* nor *short-term* fitness sacrificing), and (ii) by "fitness" we mean inclusive fitness, and (iii) by "*solving* the paradox of evolutionary altruism" we mean showing how such altruism is possible, then I see no reason at all for thinking that this frequently repeated claim is true. It is possible, however, that reciprocity is an important process by which traits regulating *helpful behaviors* evolve. For these reasons, what Trivers called instances of "reciprocal altruism" I prefer to call *reciprocal exchanges* or just *reciprocity*.[14]

Trivers thought that one way of modeling the reciprocal exchanges observed in nature is via the prisoner's dilemma—long the fascination of game theorists. The prisoner's dilemma (PD) involves two individuals who are deciding how to interact: They can both cooperate, or they can both defect, or one may offer cooperation while the other defects. But they have to make a decision simultaneously, and then compare results. Each possible outcome is associated with a "payoff" for the players (figure 1.1). In the conventional labeling, 8 is R (for *reward for cooperation*), 10 is T (for *temptation*), 1 is S (for *sucker's payoff*) and 3 is P (for *punishment for joint defection*). A prisoner's dilemma requires that T > R > P > S, and that 2R > T + S.[15]

If you are allowed to play this game just once, with one other player, it is difficult to know what to do. You might feel that mutual cooperation would be a good outcome, but to choose "cooperation" as your choice immediately opens you to exploitation. Can you trust your opponent not to leave you with 1? Better, perhaps, to be on the safe side and choose "defect," since at least getting 3 is better than 1. Of course, if the opponent reasons in the same manner, you'll both end up defecting. Things are different in an *iterated* game, when you're going to play a whole series of games with the same person, though you don't know how many games. There you need to develop a strategy which may be sensitive to what the player did on previous rounds. You may decide to defect for a while, and then try to "apologize" by offering cooperation. Or you may simply decide to always defect regardless of what your opponent does. The strategy made famous by Robert Axelrod (though it was designed by Anatol Rapoport) is known as "tit for tat" (TFT) (Axelrod 1984). TFT is terribly simple: Offer cooperation on the first round, and from then on just imitate your opponent's previous move. This amounts to cooperating so long as the opponent is cooperating,

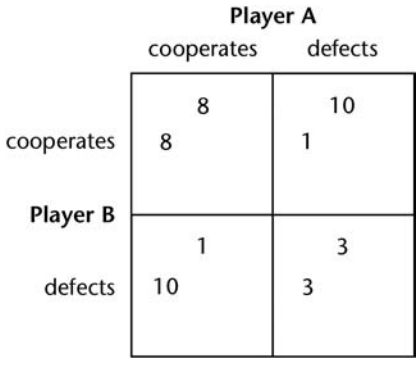


**Figure 1.1**

never defecting first, responding to any defection with your own prompt defection, and, if involved in mutual defecting, waiting patiently for the opponent to "apologize" (for she must have started it). TFT is "friendly," not open to serious exploitation, and not exploitative.

Think again of our free-riding, non-grooming monkey. Suppose she offers herself for grooming for the first time to a non-kin individual (call him "A") who promptly grooms her. Later, A offers himself to the free rider and gets nothing. Since A is "playing TFT," he will not groom her again, unless she decides at a later date to groom him. So far the free rider is up on the deal, since she got one free groom whereas A spent half an hour doing something for nothing. But if we give consideration to all the other grooming interactions going on, then she is not winning at all. She got one free groom from A, and let's say she manages to get one free groom from every other member of the group (each of whom is also playing TFT). After that, she is out of luck; no one will touch her (except kin, and we'll assume, in order to make the point, that the attentions of kin alone are insufficient to fend off parasite infestation). The others, meanwhile, are happily grooming each other for as long as they keep interacting. The non-groomer dies of parasite infestation. So much for free riding!

A common misconception is that TFT wins always, or nearly always. On the contrary, TFT never wins. The only way of ever getting more points than an opponent in a round is to defect while she offers cooperation—and by definition TFT will do this only when it has already been on the receiving end of the same treatment in an earlier round. The best TFT can do against any opponent is draw. However, TFT can win if by "winning" we mean something different. If there are a whole bunch of strategies playing off against each other (and perhaps versions of themselves), and winning consists of having the most points at the end of the whole tournament (and it is not a "knock-out" tournament), then TFT can prosper. Although with any given opponent it only draws at best, if all its opponents encounter fluctuating fortunes when playing with each other, then TFT can end up winning. Depending on the design of the tournament, it usually does remarkably well.

But in fact things are considerably more complicated than this. The triumph of TFT is entirely the result of the way the game has been set up, and there are a number of reasons for thinking that the rules of the game fail to model many aspects of real-world reciprocal exchanges. (For further discussion, see Hirshleifer and Martinez Coll 1988.) Here are half a dozen.

1. The whole point of a PD game is that players make their choices simultaneously, whereas Trivers emphasizes the time lag that characterizes recip-

rocal exchanges. Suggestion for improvement of model: Introduce the *alternating* prisoner's dilemma game, in which players know their opponent's move before they make a decision.

2. Creatures in the real world are not infallible; mistakes are likely to occur in communicating to each other. This is disastrous for two TFT players happily cooperating: If one thinks that the other has defected, it will immediately defect, leading to ongoing mutual defecting. Suggestion for improvement of model: Introduce "noise" into the game, whereby there is some probability of miscommunications and accidents.

3. In evolutionary terms, some strategies are going to be more costly to play than others. Someone playing TFT has to deploy skills of discrimination that someone playing "always cooperate" (ALL C) or "always defect" (ALL D) does not. Thus in a population of only TFT-ers and ALL C-ers—all busily cooperating with each other—those playing ALL C will have a fitness advantage, and thus will take over. Suggestion for improvement of model: Introduce a "complexity tax" on strategies.

4. We are often in a position to observe others interacting before we need to interact with them. In other words, before we sit down and make the first move with our opponent, we might have good grounds for believing what kind of a strategy she pursues. This might well affect what kind of strategy we offer her, how forgiving we are of her occasional "defect," etc. Suggestion for improvement of model: Allow players to develop a "reputation," and to alter their strategy according to the reputation of the co-player. (This may involve offering players a "scrutiny deal," such that they can gather varying degrees of information on their potential co-players at a proportional cost.)

5. In a standard PD tournament one is locked into playing with a partner regardless of how unpleasant a strategy he is employing, but in real life one can often simply choose to abandon an interaction. Combined with 4, one might choose not to interact with someone at all on the basis of his reputation. Suggestion for improvement of model: Allow *refusal to play (any more)* to be an option in the game.

6. Though TFT is said to "punish" those opponents who defect, it's really not much of a punishment. It's not even necessarily "an eye for an eye," since the only way to give the cheating opponent exactly the treatment she dealt (the "sucker's payoff") would be to *force her* to cooperate while you defect. Trivers talked of reciprocal exchanges in humans as being characterized by "moralistic aggression." This is more than just TFT's response of "Well, I'll defect with you from now on, until you mend your ways"; rather, it's a positive *penalty* of disapproval, ostracization, or possibly violence

leveled at a defector. Suggestion for improvement of model: Allow players to punish others (at a price) beyond merely defecting on them.

Most of the above features have been tried in PD tournaments (though with some of them one wonders whether it still counts as a prisoner's dilemma at all), and the conclusion is that TFT does not come out on top. First consider the introduction of "noise" into the interacting environment. As noted, this spells a disaster for the stable evolution of TFT. An alternative strategy—one that deserves its 15 minutes of fame—is called "PAVLOV." PAVLOV (whose abilities were discovered by Martin Nowak and Karl Sigmund (1993)) follows a "win-stay, lose-shift" strategy, where *winning* means receiving the R or the T payoff and *losing* means receiving the S or the P payoff. PAVLOV is far more forgiving of accidents than TFT. Suppose two PAVLOV players, Ernie and Bert, are busily engaged in mutual cooperation, when Ernie accidentally hits the *defect* button. Bert lost that round, so immediately switches to "defect" for the next round. Ernie, meanwhile, stays playing *defect*, since he won with it on the previous round. Having, then, both defected, both players immediately flip back to joint cooperation. (An exclamation mark indicates noise interference.)

Ernie: . . . C C D! D C C . . .
Bert: . . . C C C D C C . . .

Nice recovery. But see what happens when PAVLOV accidentally reveals an undiscriminating cooperator. In the first pair of rows, noise disrupts PAVLOV; in the second pair, it disrupts ALL C.

PAVLOV: . . . C C D! D D D . . .
ALL C: . . . C C C C C C . . .

PAVLOV: . . . C C C D D D . . .
ALL C: . . . C C D! C C C . . .

Ruthless exploitation, until noise interferes again. Some commentators, vaguely aware that TFT is not the end of the story in PD tournaments, nevertheless endorse the indistinct claim that "TFT-like" strategies will win the day. But if we think that TFT's "non-exploitative" characteristic is important, it is clear that PAVLOV does not count as "TFT-like." It is merciless toward the foolishly friendly, and this contributes a great deal to its success.[16]

Though organisms probably pursue something like a "win-stay, lose-shift" strategy against the environment (in making foraging decisions, for example), it would be a mistake to expect it to be an evolved strategy that dominates intelligent creatures' social interactions. Why? An important source of PAVLOV's superiority is the fact that it uses noise to weed out the

ALL C players, and then profitably exploits them to death. However, if there is a crucial advantage to be had from uncovering the undiscriminating players and exploiting them, it is not plausible that biological natural selection would plump for the inefficient mechanism: "Wait until you accidentally defect, then see what happens." If such a value is to be had from uncovering the undiscriminating, then natural selection is likely to prefer a somewhat more direct means of flushing them out. One might instead try *purposely* defecting to see what happens. But such an experimental defection might meet with a severe penalty. (If you're wondering whether a country's laws uphold the death penalty for treason, an especially poor way of satisfying your curiosity would be to travel to that country, commit treason, and see what happens.) If identifying suckers and defectors is important, then probably the best way to do it is to observe other players interacting. Needless to say, gathering information may cost something (in fitness terms), but the rewards of having advance warning about what kind of strategy your partner is likely to deploy may be considerable. There have been several attempts to model this element in PD playoffs (e.g., Sugden 1986; Pollock and Dugatkin 1992; Nowak and Sigmund 1998; Panchanathan and Boyd 2003, 2004). Usually, however, the notion of reputation that is employed reflects only whether one has given unprompted defections in the past. But the success of PAVLOV suggests that reputations should also reflect whether one is an unconditional cooperator.

## 1.5 The Evolution of Helping: Indirect Reciprocity

"The purest treasure mortal times afford," Shakespeare tells us, "is spotless reputation." In his less flamboyant manner, Darwin agreed: ". . . love of praise and the strong feeling of glory, and the still stronger horror of scorn and infamy" are together a "powerful stimulus to the development of the social virtues" ((1879) 2004: 133, 156). By introducing reputation into our understanding, we move away from standard reciprocal exchanges to what has been called "indirect reciprocity." This lies at the heart of Alexander's account (1987) of the evolution of moral systems, and I agree that it is of central importance. In indirect reciprocal exchanges, an organism benefits from helping another by being paid back a benefit of greater value than the cost of her initial helping, but not necessarily by the recipient of the help. We can see that reputations involve indirect reciprocity by considering the following: Suppose A acts generously toward several conspecifics, and this is observed or heard about by C. Meanwhile, C also learns of B acting disreputably toward others. On the basis of these observations—on the basis, that

is, of A's and B's reputations—C chooses A over B as a partner in a mutually beneficial exchange relationship. A's costly helpfulness has thus been rewarded with concrete benefits, but not by those individuals to whom he was helpful. Alexander lists three major forms of indirect reciprocity:

> (1) the beneficent individual may later be engaged in profitable reciprocal interactions by individuals who have observed his behavior in directly reciprocal relations and judged him to be a potentially rewarding interactant (his "reputation" or "status" is enhanced, to his ultimate benefit); (2) the beneficent individual may be rewarded with direct compensation from all or part of the group (such as with money or a medal or social elevation as a hero) which, in turn, increases his likelihood of (and that of his relatives) receiving additional perquisites; or (3) the beneficent individual may be rewarded by simply having the success of the group within which he behaved beneficently contribute to the success of his own descendants and collateral relatives. (1987: 94)

One possible example of indirect reciprocity is the behavior of Arabian babblers, as studied by Amotz Zahavi over many years (Zahavi and Zahavi 1997). Babblers are social birds that act in helpful ways toward each other: feeding others, acting as sentinels, etc. What struck Zahavi was not this helpful behavior per se, but the fact that certain babblers seem positively eager to help: jostling to act as sentinel, thrusting food upon unwilling recipients. The "Handicap Principle" that Zahavi developed states that such individuals are attempting to raise their own prestige within the group: signaling "Look at me; I'm so strong and confident that I can afford such extravagant sacrifices!" Such displays of robust health are likely to attract the attention of potential mates while deterring rivals, and thus such behavior is, appearances notwithstanding, squarely in the fitness-advancing camp.

Consider the enormous and cumbersome affair that is the peacock's tail. Its existence poses a prima facie threat to the theory of natural selection—so much so that Charles Darwin once admitted that the sight of a feather from a peacock's tail made him "sick!" (F. Darwin 1887: 296). Yet Darwin also largely solved the problem by realizing that the primary selective force involved in the development of the peacock's tail is the peahen's choosiness in picking a mate.[17] If peahens prefer mates with big fan-shaped tails, then eventually peacocks will have big fan-shaped tails; if peahens prefer mates with triple-crested, spiraling, red, white, and blue tails, then (ceteris paribus) eventually peacocks will sport just such tails. Sexual selection is a process whereby the choosiness of mates or the competition among rivals can produce traits that would otherwise be detrimental to their bearer.[18] I am not categorizing sexual selection in general as reciprocity, only those examples

that involve the favoring of traits of costly helpfulness. If a male is helpful to a female (bringing her food, etc.) and as a result she confers on him the proportionally greater benefit of reproduction, this is an example of direct reciprocity. If a male is helpful to his fellows in general, and as a result an observant female confers on him the proportionally greater benefit of reproduction (thus producing sons who are generally helpful and daughters who have a preference for helpful males), this is an example of indirect reciprocity.[19] Just as sexual selection can produce extremely cumbersome physical traits, like the peacock's tail, so too can it produce extremely costly helping behaviors. We can say the same of reputation in general if the benefits of a good reputation are great enough. If a good reputation means sharing food indiscriminately with the group, then an indiscriminate food-sharing trait will develop; if a good reputation means wearing a pumpkin on your head, then a pumpkin-wearing trait will develop. The same, moreover, can be said of punishment, which is, after all, the flip side of being rewarded for a good reputation. If a type of self-advancing behavior (or any type of behavior at all) is sufficiently punished, it will no longer be self-advancing at all. (See Boyd and Richerson 1992.)

Once we see that indirect reciprocity encompasses systems involving reputation and punishment, and that these pressures can lead to the development of just about any trait—extremely costly indiscriminate helpfulness included—then we recognize what a potentially important explanatory framework it is. As a way of reminding ourselves of how important reciprocity can be, we should recall Aristotle's shrewd observation in *Politics* that for creatures who trade there is nothing that has only one function: A spear is good for hunting but may also be swapped; a pot is handy for carrying water but may also be used to bargain with; the skill of gathering foodstuffs contributes to satisfying nutritional needs but may also be exchanged for other favors; and so on. That the advantages of doubling the functionality of one's resources are considerable is obvious.

## 1.6 The Evolution of Helping: Group Selection

Skills of discrimination lie at the heart of both direct and indirect reciprocal exchanges. In the former, one helps only those who will help one back; in the latter, one favors or punishes others depending on their past performance. But there are other models to which we can appeal that need involve no such powers of discrimination on the part of the helpers. These helpers need not be reciprocal helpers at all; they will help anyone at all in their group, irrespective of the treatment they receive in return. It seems

hardly credible that natural selection could favor such behavior. One way of putting this incredulity is to note that such helpers would appear to be genuine *fitness sacrificers*. But how could natural selection possibly favor a fitness-sacrificing creature over a fitness-advancing creature?

In their defense of group (or multi-level) selection, Elliott Sober and David Sloan Wilson (1998) have shown how it can work. First let's give a model in which helping *doesn't* take off. Suppose we have a population of 200 individuals. It doesn't matter whether they're humans, frogs, lions, plants, or computer programs. The important point is that they reproduce at a certain rate. Let's put the base-line rate at 1.1. By "base-line rate" we mean the rate at which an individual reproduces without any interference: without receiving any help, without making any sacrifices. So if there were nothing special going on in the population—no sharing or sacrificing—then it would grow by 10 percent each generation. (To make things simpler, we're assuming that reproduction is asexual and that the old generation immediately disappears upon the arrival of the new.) But let's put into the midst of this population a few helpful individuals. In performing helpful behaviors toward non-kin, helpers lower their own reproductive capacity. Let's say that they sacrifice 9 percent of their reproductive capacity in order to help 10 of their comrades get a boost of 0.4—that distribution of benefits being uniform and undiscriminating of whether the recipient is helper or non-helper. If we put 10 such characters into the mix, then 100 individuals will get their capacity boosted by 0.4. And since the distribution is uniform, five of those 100 will be helpers. We end up with the following spread of reproductive capacities in our population: five helpers with a reproductive rate of 1.4, five with a rate of 1.0, 95 unhelpful individuals with a rate of 1.5, and 95 on the base line with 1.1. This results in the following for generation 2:

population = 259 (12 helpers, 247 non-helpers).

First note the enormous impact that the helpers have had. Instead of the base-line growth of 10 percent, the total population has grown by almost 30 percent. But note secondly the percentage of helpers in the new population: It has dropped from 5 percent to 4.6 percent. After another round, things look as follows for generation 3:

population = 330 (14 helpers, 316 non-helpers).

The percentage of helpful individuals has dropped further, to 4.2. And if we carry on we'll see it continue to drop. If there's any environmental pressure restricting population growth—as there must be—then helpfulness, for all

the benefits it has brought to the group, goes extinct. Not only does helpfulness have trouble getting established in a population, but it's vulnerable to overthrow. If we run a similar test, but this time starting with 199 helpers and just one non-helper, then despite the fact that the group will grow explosively with the values assigned, gradually, steadily, the percentage of unhelpful individuals in the population will increase, and the helping trait is doomed.

Now let's turn to group selection. Start again with 200 individuals, 10 of whom are helpers, just as we did above, with the same values holding. But this time the population is split into two groups of 100 each, which are, for a time at least, isolated. One of the groups—group A—has only unhelpful individuals therein, so will grow at a rate of 1.1, increasing to 110 individuals in the second generation. Group B has the 10 helpers, who give out the same benefit as above (a boost of 0.4 to 10 fellows, spread evenly and indiscriminately). In the second generation, group B will have 149 individuals, 14 of whom are helpers. The interesting thing is that the percentage of helpers relative to the size of group B has fallen from 10 percent to 9.4, whereas relative to the global population (A + B) their percentage has risen from 5 to 5.4. (This is an instance of Simpson's Paradox; see Simpson 1951.) After another round, things look as follows for generation 3:

| Group A | Group B |
|---|---|
| population = 121 | population = 218 |
| (all non-helpers) | (19 helpers, 199 non-helpers). |

The percentage of helpers relative to group B has dropped further (to 8.7), while rising further relative to the global population (to 5.6). If we just went on like this, nothing interesting would have been shown; assuming a limit on population growth, we would observe the trait of helping run to extinction just as we did before. But suppose that before this occurs the population is shaken up in some special way. Imagine that the total population—both groups A and B, now standing at 339—is mixed together and proportionally cut back to its original size of 200, again in two groups. Since the percentage of helpers will have grown in the interim, they will enjoy a larger representation in the new starting lineup than they began with. And here's the important part: Suppose we allow members to express some preference regarding with whom they associate in these new groups (such that no one is able to force unwanted association upon another). This preference might be nothing more than selecting individuals from whom one is likely to gain. Anyone stands to gain most from getting into a group with helpers, and this includes helpers themselves. The consequence will be

a tendency for the helpers to "clump" together when the two new groups form.[20] In the third generation the percentage of helpers had reached 5.6, which out of 200 amounts to 11. Suppose all these helpers clump together in one of the groups of the new starting lineup.

So now we start again, with two new groups: A and B. Everything is the same as before, except that now group B begins with 11 helpers instead of 10. If we run it again to the third generation before shaking things up, then the percentage of helpers, relative to the total population, is over 6. If we cut back again to two groups of 100, assuming again that the helpers end up clumping, then this time there will be 12 helpers in group B's starting lineup. If we run it to the third generation one more time, then helpers reach over 7 percent of the global population, putting even more into the new starting lineup. Things are starting slowly, but if we were to go on running this growth program we would see a curve favoring the takeover of the helping trait. I purposely haven't made things easy for the helpers. If we allowed them the option of sacrificing a further 0.1 of their reproductive capacity in order to help another 10 individuals each get a bonus of 0.4, then their numbers would take off much faster. By comparison, in the "single-group model" that we considered above—where all 200 individuals were lumped into one undifferentiated group—this further sacrifice on the part of the helpers would just have led to their swifter demise.

We saw how, in the single-group model, a population of non-discriminating helpers was terribly vulnerable to takeover by non-helping individuals. What happens in this multi-group model when non-helpers turn up in a population of helpers? Let's run a similar model to the previous one, but starting out with 10 non-helpers in group B (not forgetting that just *one* was enough to take over in the single-group model). Both groups A and B are otherwise populated entirely by helpers. Here's how things go.

Generation 2:

| Group A | Group B |
|---|---|
| population = 500 | population = 461 |
| (all helpers) | (414 helpers, 47 non-helpers) |

Generation 3:

| Group A | Group B |
|---|---|
| population = 2,500 | population = 2,125 |
| (all helpers) | (1,904 helpers, 221 non-helpers) |

Helping, as we saw above, can potentially allow for remarkable growth. At a glance, the trait of unhelpfulness appears to be doing well: Its numbers have jumped from 10 to 47 to 221. Indeed, it is slowly taking over group B.

But notice what has happened to the percentage of non-helpers in the global population. It has dropped from 5 to 4.9 to 4.8. Suppose we were then, as before, to pare the whole lot back to two groups of 100 each (preserving the ratio of helpers and non-helpers when we do so), allowing individuals to choose with whom they associate. There will be fewer non-helpers to go into the mix. Again, everyone—both non-helpers and other helpers—wants to be with helpers, and so non-helpers get consigned to a group by themselves. If we run it out again to the third generation the percentage of non-helpers will have dropped even further, putting even fewer offspring into the next starting lineup. And so on, till unhelpfulness goes extinct.

One might complain that this is all just fiddling with numbers to get the desired result. There is a hint of truth in that. But remember that the objective is to show how helping behavior *could* develop through the forces of biological natural selection. Natural selection could spend millions of years throwing up a whole range of characters who help too much, or not enough, or in the wrong way, and on all such occasions the trait falls flat. But if among the myriad of values that won't work there is one that strikes the right balance and allows helping behavior to take off, then we might expect it to be eventually hit upon. Nature is nothing if not patient.

Nor should this business of allowing things to run for three generations and then shaking them up be taken literally. Waiting for three generations is just a useful illustrative means of showing how the frequency of traits can grow or fall relative to different domains. It is not being suggested that any actual population follows this "three generations, followed by regrouping" pattern. The important point is the tendency of the helpers to associate together, which follows directly from the dictum "Everyone loves a helper." Groups containing helpers will outperform groups containing fewer or no helpers. And thus the helping trait—that is, indiscriminate, non-reciprocating helping—can develop. It is pedagogically useful to think of the multi-group model as involving, say, tribes in neighboring valleys, or mice living in separate haystacks, all of whom periodically come together to mate and form new groups, but it is just that: a useful picture. It can work just as easily in a population that to all appearances is one big group, so long as we allow that within that group the helpers are tending to associate together.

Are the helpers being described here really sacrificing their fitness? That would be the equivalent of genuinely "altruistic genes"—to stand in contrast to Richard Dawkins's famous metaphor. We saw that in the case of direct reciprocal helping the fitness sacrifice was only apparent. The big fish of the coral reef sacrifices a free meal, but only because it gets a valuable

long-term payback (ongoing freedom from parasites). The helpers we are now considering really do seem to be giving up some reproductive fitness: They have a trait of advancing the reproductive interests of others at the expense of their own fitness. In the original group B (90 non-helpers with a fitness of 1.1, plus 10 helpers with a fitness of 1.0), who has the higher starting fitness? The non-helpers. Now let's take the benefits that those 10 helpers have to offer and distribute them among the group. On average, the helpers' fitness will rise to 1.4, while the non-helpers' will rise to 1.5. It still pays to be unhelpful.

The only way to calculate the numbers such that being a helper actually turns out to *increase* the individual's fitness is to include group A in our figuring. When we look at the fitness of the average helper across *both* groups, then we see that it is 1.4 compared to the non-helpers' average fitness of 1.28. But this, it has been argued, is not the pertinent calculation. Sober and Wilson refer to this as an instance of "the averaging fallacy." Their main concern is that focusing on the global calculation in order to determine fitness obscures the dynamics of the processes that are really at the heart of selection. If we ignore the fact that there are two groups growing at different rates, and instead just said there is one big group of 200 individuals, 10 of whom have a fitness of 1.4 and 190 of whom have a fitness of 1.28, we have lost sight of what *explains* these figures. What explains them is that the former 10 individuals are distributing a benefit that is available to 90 of the latter individuals (plus themselves) and unavailable to the other 100.

It seems that the only judicious conclusions are (1) that in one sense these helpers are genuinely fitness sacrificing but in another sense they aren't and (2) that there is an argument is favor of the greater explanatory productiveness of the former perspective. I am willing to end with this somewhat uncommitted view on the matter, since the existence of genuinely fitness-sacrificing traits is not necessary to this project. What is important is that *helping* behaviors have been selected for, and the multi-group approach provides a further model for how that might occur.

We have now seen how this multi-group model might allow non-discriminating helpfulness to develop. Clearly, if we were to run the same multi-group model with *discriminating* helpfulness, then helpful behavior would develop all the more easily. The values assigned to helpfulness can be such that the dynamics of a group-structured population alone will be insufficient for its development, whereas if we make the helpfulness a little discriminating (e.g., such individuals are reluctant to help the unhelpful) then the trait will evolve. In other words, there exist circumstances in which group selection and reciprocity together may lead to a degree of help-

fulness that either process alone could not produce. We need not see them as exclusive alternatives.

There is a stronger sense in which reciprocity and group selection might not be alternatives: namely, if reciprocity just *is* a form of group selection. Sober and Wilson would object, I think, to the way I have structured the preceding few sections, since they would argue that kin selection, mutualism, and reciprocity should all be subsumed under group selection. Before moving on, I should briefly say something to defend my taxonomy on this score.

Consider direct reciprocal altruism. Sober and Wilson argue that the relevant notion of a group constituting a vehicle of selection is a *trait group*—a population of $n$ individuals (where $n > 1$) "that influence each other's fitness with respect to a certain trait but not the fitness of those outside the group" (1998: 92ff.). On this basis, they conclude that reciprocal altruism is really just a special form of group selection, involving a group of two. But Kim Sterelny (1996) has argued plausibly that there is a difference *in kind* between groups that satisfy the above criterion (including partners in reciprocal exchanges) and the "superorganisms" often used as paradigmatic examples of group selection (including especially colonies of social insects). Examples of the latter category exhibit an extreme degree of cohesion and integration, their members share a common fate, and such groups possess adaptations that cannot be equivalently re-described at the individual level (e.g., the tendency of newly hatched queens to kill their sisters). Such groups have as respectable a claim to being robustly objective vehicles of selection as do organisms. Concerning examples of the former category, by contrast, the decision to describe selection as occurring at the level of the group is a purely optional one, for this group-level description is equivalent to an individual-level description. Regarding this category, Sterelny (following Dugatkin and Reeve 1994) advocates a pluralistic approach, where the only difference between preferring individuals or trait groups as the vehicle of selection—that is, of regarding the process as one of individual selection or group selection—is a heuristic one, depending "on our explanatory and predictive interests" (1996: 572).

Going along with Sterelny, I am willing to concede that, on a certain liberal understanding of what it takes to be a group, reciprocal relations may count as group selected, or they can be equivalently described in terms of individual selection. Any debate on the matter, says John Maynard Smith, is not "about what the world is like . . . [but] is largely semantic, and could not be settled by observation" (1998: 639). But it is clear that there is a kind of group selective process which they are *not* an example of: what Sterelny calls "superorganism selection" (1996: 577). One could argue that human

cooperative faculties (e.g., morality) are the product of superorganism selection, or one might instead argue that they may be explained by invoking only, say, reciprocity. These are quite distinct hypotheses, and it cannot be reasonably denied that if we were unable to distinguish between them due to a methodological decision to lump reciprocity (along with kin selection) under the umbrella term of "group selection" this would be an unacceptable loss of explanatory detail in the service of theoretic unification.

## 1.7 The Evolution of Human Ultra-Sociality

I have reviewed four processes whereby traits of helpfulness can develop by the forces of biological natural selection: kin selection, mutualism, reciprocal exchanges (both direct and indirect), and group selection. The biologist Lee Dugatkin (1999) has called these "the four paths to cooperation"; they are almost certainly the most important processes by which traits of helpfulness evolve in the animal world, though we should be open to the possibility of others (e.g., Connor 1995; see also Sachs et al. 2004). However, it is not at all clear that these processes alone can account for the ultra-sociality that is characteristic of human life. It is not unreasonable to view human social complexity as having more in common with the cooperative life of social insects than with the small-scale groupishness of our closest primate cousins. Yet unlike the Hymenoptera, whose traits of extreme helpfulness appear to be due to unusual genetic relations, an important part of the explanation of human ultra-sociality is surely our unique capacity to transmit masses of adaptive cultural information in a cumulative way. Though the kinds of reciprocity I have discussed almost certainly have played a major role in human ancestry, and have left their marks on the human mind, available models (Boyd and Richerson 1988, 1989, 1992) suggest that they will work only for relatively small groups: something along the lines of a chimpanzee troop. This is not a problem for the hypothesis of this book, for it is quite possible that morality evolved when our ancestors were still in relatively small bands. However, insofar as this chapter has the more general aim of outlining processes that can lead to helpfulness, any discussion is incomplete to the extent that it fails to explain human ultra-sociality.

The apparent fact that reciprocity alone cannot explain large-scale helpfulness isn't altered even if we factor in the possibility of punishing non-reciprocators, since doing so leaves us with the question of why those administering the punishments don't lose out in the evolutionary struggle to "easy-going reciprocators": those who reciprocate but aren't willing to

expend energy on punishing others. Sober and Wilson appeal to group selection to explain how punishing traits evolve: A group of punishers may outperform a group of easy-going reciprocators. One might wonder why their model need invoke punishment at all. Couldn't group selection just directly produce reciprocal helpfulness? Answer: It *could*, but it is much more likely that group selection will produce helpfulness via punishing traits than that it will produce them directly. The reason for this is that, although administering punishment generally costs the administrator something, typically it doesn't cost her much (proportional to the group benefit provided). Suppose you own a small business and someone comes in one day and asks you to give him 20 percent of your monthly earnings. That's quite a sacrifice. But if the penalty of forfeiture is death, then handing over the 20 percent is the prudent thing to do. Now compare your sacrifice with how much it costs the racketeer to create a credible threat of penalty. Perhaps all he need do is occasionally drive slowly past your house in a menacing manner.

When a fitness-sacrificing trait evolves by group selection, it is always as a result of winning a competition: The forces of individual fitness advancement tug in one direction, the forces of group-benefiting fitness sacrifice in the other. A major contribution of Sober and Wilson's work is to show how the former forces need not always win. But obviously the fitness-sacrificing forces are more likely to win when the fitness-advancing forces are smaller; in other words, behaviors that benefit the group but cost the individual a great deal are less likely to evolve than comparable traits that cost the individual less. And if the punishment is the withdrawal of social esteem (McAdams 1997), which can be distributed or denied like a magical substance, or exclusion from ongoing beneficial exchanges (Panchanathan and Boyd 2004), then punishment can often be meted out at no cost.

However, though no one believes that genetic group selection is impossible, it is questionable how large a role it played in human ancestry. The main hindrance is the degree to which group membership affects mating choices. Two tribes of humans may be in intense competition, but any allowance of intermarriage or migration between the tribes will count against the likelihood that group selection is taking place at the genetic level. Even if the two tribes participate in all-out warfare, so long as the result of victory is the taking of the women of the conquered tribe, or the assimilation of the survivors, then genetic selection is militated against.

But group selection need not occur at the genetic level. Bear in mind that natural selection is not concerned essentially with genes at all. Darwin articulated the theory beautifully while remaining utterly ignorant of

genetics. So long as there is trait variation, heritability, and trait-dependent differential reproduction, then there is selection. (See Lewontin 1970.) (If this selection is guided by purposeful design, it is artificial selection; otherwise it is natural selection.) There is nothing in the theory that says that the traits in question must be genetically encoded, or that the reproducing entities must be individual organisms. Learned cultural practices may result in trait variation among groups, may be transmitted between groups, and may affect the persistence and proliferation of groups; thus *cultural* group selection may occur in circumstances that are not conducive to genetic group selection.[21]

In order for group selection to occur, there must be a degree of uniformity within groups and a degree of variability between groups. Though it is not impossible that these criteria may be satisfied at the genetic level, such widespread phenomena as migration and intermarriage present obstacles to their actual satisfaction. These criteria seem much more plausibly satisfied at the cultural level. The anthropologists Joe Henrich and Robert Boyd (1998) show how having a tendency to conform one's behavior to that of the majority of one's group can be adaptive in a variable environment, since it allows reliable and efficient access to those behaviors that are likely to be successful in the immediate environment. (See also Boyd and Richerson 1985.) Copying the successful, or (what will tend to amount to the same thing) copying the majority, can allow individuals to "short-cut the costs of individual learning and experimentation, and leapfrog directly to adaptive behaviors" (Henrich and Boyd 2001: 80). Thus, Henrich and Boyd hypothesize that genetic evolution has produced in humans psychological mechanisms that support conformist transmission, and, further, that this trait lies at the heart of humans' unique cumulative culture. Especially when coupled with traits pertaining to the employment of punishment strategies, conformist transmission explains how within-group cultural variation may be suppressed while intergroup variation is enhanced. The thing about punishment, as we saw earlier, is that it can in principle fix just about any behavior in a group, even weird and seemingly maladaptive behaviors. But this is where cultural group selection may play an important role: Once there exists a meta-population of culturally distinct groups, there is selective pressure in favor of the persistence and proliferation of those cultural traits that are broadly "prosocial." A group whose cultural value system revolved around wearing a pumpkin on one's head would, on the whole and in the long run, lose out to a group that valued intragroup peacefulness and a degree of self-sacrifice for the welfare of one's fellows. This, then, is another theory explaining the evolution of helpfulness.

An important addition to this story of cultural group selection is that in creating cultures our ancestors enormously influenced the environmental niche within which they lived. There is no reason to doubt that as cultural group selection occurs genetic individual selection may be ongoing, and if this has been the case then the course of the latter process will have been highly influenced by the outcome of the former process. For instance, a cultural activity such as dairy farming may affect the genetic makeup of the population by favoring the trait of lactose tolerance. In West Africa the cultivation of yams led to the clearing of the rain forest, which resulted in more standing water, which allowed more mosquitoes to breed, and the consequent multiplication of the malarial risk magnifies the pressure in favor of the sickle-cell allele, which in its heterozygotic form gives protection against malaria (Durham 1991). It has even been hypothesized that the ancient cultural invention of cooking meant that less energy had to be expended on the human digestive system, making possible the explosive growth of the energy-hungry hominid brain (Aiello and Wheeler 1995). If this is correct, then it is not only true that our big brains made possible culture; it is also true that that culture made possible big brains.

In much the same way, if we allow that cultural group selection can produce a climate within which non-reciprocation will be punished, and perhaps also where a reluctance to punish non-reciprocators will be punished, and perhaps also where non-conformity to the majority will be punished, then we must allow that individuals within this environment may have new selective pressures upon them—pressures that did not exist before cultural evolution. Thus individual selection occurring at the genetic level could now produce psychological traits designed to enhance one's success in this environment where prosociality is so heavily rewarded. (See Henrich and Boyd 2001.) Individual genetic evolution and cultural group evolution may then engage in a positive feedback loop, producing not only highly social creatures but creatures whose ultra-sociality is to a significant extent genetically encoded.[22] Since humans are the only known organisms for whom significant cultural evolution occurs, this process of cultural-genetic coevolution is a special case among the explanations of animal helpfulness.

## 1.8 Conclusion and Preview

A great deal more could be said about all these evolutionary processes that favor the development of helpfulness, but a more detailed taxonomy is not my concern here. The question to which I now turn concerns natural selection's

*means* of achieving helpfulness. Suppose that in a population of ancestral bees there is pressure in favor of additional helpfulness, and the explanation of this pressure is kin selection. Knowing that this is the source of the pressure doesn't tell us anything about *how* the additional helpfulness might be achieved. One thing we know is that natural selection can't achieve the result of a more helpful bee by magic; it must go to work on whatever mechanisms are already in place governing the organism's behavior, tweaking them or transforming them so as to encourage new or stronger helpful behaviors. For this reason, there is no general answer to the question of what means Mother Nature employs in order to achieve helpfulness, any more than there is one concerning the means by which organisms achieve locomotion. The mechanisms in place that determine the helping behaviors of bees are unlikely to bear much resemblance to those that ensure the helping behaviors of chimpanzees. The evolutionary *processes* that explain such helpful behaviors may be broadly the same (it may be kin selection in both cases, for example), but the *means* by which those processes achieve results are going to differ remarkably.

The thesis to be examined in the next three chapters is that among the means favored by natural selection in order to get humans helping each other is a "moral sense," by which I mean a faculty for making moral judgments. It is possible that such a mechanism may be the result of any of the processes outlined above, or any combination of them. My own judgment is that if there is such an innate faculty the process that most probably lies behind its emergence is indirect reciprocity, but it is not an objective of this book to advocate this hypothesis with any conviction. (I do, however, discuss the matter further in section 4.6. See also Joyce forthcoming c.) We have a prior and more pressing task to attend to, which will require us temporarily to put aside issues pertaining to evolution. Any attempt to understand how our ability to make moral judgments evolved will not get far if we lack a secure understanding of what a moral judgment is. (To neglect this would be like writing a book called *The Origins of Virtue* without any substantial discussion of what virtue is.) This is the purpose of the next chapter—though, for reasons that will unfold, the chapter will start out discussing kin selection and love.