CHAPTER 1

The Springs of Sympathy

§3. Psychological Altruism: Basics

At some point in our evolutionary past, before the hominid line split off from the branch that leads to contemporary chimpanzees and bonobos (possibly quite a long time before), our ancestors acquired an ability to live together in small groups mixed in terms of sex and age. That achievement required a capacity for altruism. It also prepared the way for unprecedented forms of cooperation, and ultimately for the enunciation of socially shared norms and the beginnings of ethical practice. Altruism is not the whole story about ethics, but it is an important part of it.[1]

My analytical history of the ethical project thus begins with a hypothesis about the social groups in which the project originated and about the

1. There is a long tradition, stemming from Hume, Adam Smith, and Schopenhauer, that places a capacity for sympathy at the center of ethics. In recent years, that tradition has been renewed by philosophers (Simon Blackburn, *Ruling Passions* [New York: Oxford University Press, 1999]) and by primatologists (Frans de Waal, *Primates and Philosophers* [Princeton, NJ: Princeton University Press, 2007]). Although the approach I shall defend overlaps with some of the themes of this tradition, it does not ascribe sympathy (or altruism) so dominant a role. For explicit comparisons, see my discussion of de Waal, "Ethics and Evolution: How to Get Here from There," in *Primates and Philosophers*.

psychological capacities of the members of those groups. Fossil evidence, together with the remains found at hominid and early human sites, reveals that our ancestors lived in bands akin to those in which chimpanzees and bonobos live today: the members were young and old, male and female; the band size was (roughly) 30–70.[2] This chapter argues that, to live in this way, hominids and human beings had to have a capacity for altruism, one contemporary people almost certainly retain. To understand the historical unfolding of ethics we shall need to recognize the intricacies of the notion, as well as the varieties and limitations of hominid/human altruism. The next sections supply the necessary preliminaries.

It is important to distinguish three types of altruism. An organism *A* is *biologically altruistic* toward a beneficiary *B* just in case *A* acts in ways that decrease its own reproductive success and increase the reproductive success of *B*. For a century after Darwin, there was a deep puzzle about how biological altruism is possible. During the past fifty years, however, that puzzle has been solved. Biologically altruistic actions directed toward kin can promote the spread of the underlying genes. Moreover, when organisms interact with one another repeatedly, biological altruism exhibited on some occasions can gain dividends from future reciprocation.[3]

2. Different anthropologists use different methods for estimating hominid group size, some favoring direct comparisons with social groups in other species (either evolutionary relatives or primates with a similar ecology), others taking extant hunter-gatherer bands as models or seeking correlations with measurable anatomical features (e.g., skull size) and extrapolating from the results on hominid skulls (viewed as providing clues to the relative increase in neocortex size). See Robin Dunbar, *Grooming, Gossip, and the Origins of Language* (London: Faber, 1996); Steven Mithen *Pre-History of the Mind* (London: Thames and Hudson, 1996); Christoph Boehm, *Hierarchy in the Forest* (Cambridge, MA: Harvard University Press, 1999); Clive Gamble, *The Palaeolithic Societies of Europe* (Oxford, UK: Cambridge University Press, 1999); and Peter MacNeilage, *The Origin of Speech* (Oxford University Press, 2008). Although I am inclined to accept a relatively small value (30–70), my conclusions would not be greatly affected were this increased to, say, 80–140.

3. The original papers are W. D. Hamilton, "Genetical Evolution of Social Behavior," I, II, *Journal of Theoretical Biology* 7 (1964): 1–52; Robert Trivers, "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology* 46 (1971): 35–57; Robert Axelrod and William Hamilton, "The Evolution of Cooperation," *Science* 211 (1981): 1390–96. Lucid and accessible summaries are available in Richard Dawkins, *The Selfish Gene*, 2nd ed. (New York:

Biological altruism requires no perceptive or cognitive abilities. Even plants can have traits that make them biologically altruistic, for their propensities to form roots or to set seeds can limit individual reproductive success and facilitate the reproduction of neighbors. For animals capable of recognizing the wishes of those around them, however, we can develop a useful behavioral analog of the notion of biological altruism.[4] An animal *A* is *behaviorally altruistic* toward a beneficiary *B* just in case *A* acts in ways that detract from its fulfillment of its own current desires and that promote the perceived wishes of *B*.[5] Behavioral altruists do what they take the animals around them to want. They may act in this way not out of any particular concern for those other animals, but because they think that some of their own wishes will ultimately be well served by doing as they do. Behavioral altruism may be practiced by Machiavellian egoists (and, as we shall eventually see—§11—it can also be practiced by individuals who fall into a category intermediate between egoism and psychological altruism).

Neither biological altruism nor behavioral altruism is of much help in understanding the origins of the ethical project. For our purposes, the significant notion is that of *psychological altruism.* Psychological altruism has everything to do with the intentions of the agent and nothing to do with the spread of genes, or even the successful satisfaction of the wishes of others. Assuming for the moment that there have been human beings who are psychological altruists, the vast majority of them have not known much about heredity, and even those who have were rarely concerned with spreading genes. They acted to promote what they took to be the wishes, or the interests, of other people.[6] Sometimes they succeeded. Yet, even when they did not, their serious efforts to do so qualified them as psychological altruists.

Oxford University Press, 1993); and Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984). I shall be exploring these important ideas in §8.

4. For discussions about behavioral altruism, I am indebted to Christine Clavien.

5. There are complications that I glide over here and that will be addressed more thoroughly in treating the third type of altruism, the one pertinent to the examination of ethics. After the presentation of that third notion, it will be easier to see how to characterize behavioral altruism more exactly.

6. As the specification of psychological altruism will show, the account begins with wishes. Interests come later (§21).

Many people believe psychological altruism does not exist, even that it is impossible. Often they are moved by a very simple line of reasoning: when a person acts in a way that could be appraised as altruistic, he or she acts intentionally; to act intentionally is to identify an outcome one wants and to attempt to realize that outcome; hence, any potential altruist is trying to get what he or she wants; but to strive for what you want is egoistic; consequently, the potential altruist turns out to be an egoist after all. The key to rebutting this argument is to distinguish different kinds of wants and goals. Some of our desires are directed toward ourselves and our own well-being; other desires may be directed toward the welfare of other people. Desires of the former type are the hallmark of egoism, but those of the latter sort are altruistic. So altruists are intentional agents whose effective desires are other-directed.[7]

I shall develop this approach to psychological altruism further, by giving a more detailed account of the character of other-directed desires, and thereby bringing into the open some of the complexities of the concept of altruism. In focusing on desires, I ignore for the moment the fact that there are other psychological attitudes—hopes, aspirations, and particularly emotions—that can be properly characterized as altruistic. Attention to these other types of states will occupy us in the next section. Because of the connection of desires with intentions and actions, altruistic desires have a certain priority. They are thus the topic of the basic account.

The other-directed desires central to the defense of the possibility of altruism are desires that respond to the altruistic agent's recognition of the impact of his or her actions on the situations of others. To be an altruist is to have a particular kind of relational structure in your psychological life—when you come to see that what you do will affect other people, the wants you have, the emotions you feel, the intentions you form, change from what they would have been in the absence of that recognition. Because you see the consequences for others of what you envisage doing, the psychological attitudes you adopt are different. You are moved by the

7. This line of response surfaces in the eighteenth century in the famous series of sermons given by Joseph Butler at the Rolls Chapel. Many subsequent writers have followed Butler's lead—as shall I.

perceived impact on someone else. If your response leads you to *act* altruistically, that is because your *desires* have been affected.[8]

So far, that is still abstract and vague. I shall motivate the underlying idea with a simple and stylized example and then offer a more precise definition.

Imagine that you are hungry and that you enter a room in which some food is spread out on a table. Suppose further that there is nobody in the vicinity who might also be hungry and want all or part of the food. Under these circumstances, you want to eat the food; indeed, you want all of it. If the circumstances were slightly different, however, if there were another hungry person in the room or believed to be in the neighborhood, your desire would be different: now you would prefer the outcome where you share the food with the other person. Here your desire responds to your perception of the needs and wants of someone else, so that you adjust what you might otherwise have wanted to align your desire with the wants you take the other person to have.

This is a start, but it is not sufficient to make you an altruist. For you might have formed the new want when you see that someone else will be affected by what you do, because you saw profitable future opportunities for accommodating this other person. Maybe you envisage a series of occasions on which you and your fellow will find yourselves hungry in food-containing rooms. You see the advantages of not fighting and of not simply having all the food go to the first person who enters. You resolve to share, then, because a future of cooperation will be better from your point of view. For real altruism, the adjustment of desires must not be produced by this kind of self-interested calculation.

I offer a definition of "*A* acts psychologically altruistically towards *B* in *C*"—where *A* is the agent, *B* is the beneficiary, and *C* is the context (or set of circumstances). The first notion we need is that of two situations differing from each other in the recognizable consequences for others (people or nonhuman animals). Let us say, then, that two contexts *C* and

8. You might be affected by another person's predicament, and form an altruistic emotion, but that might not generate a desire that issues in action. The most basic type of altruism that is of ethical concern is a response to someone else that eventually expresses itself in conduct.

*C** are *counterparts,* just in case they differ only in that, in one (*C**, say) the actions available to *A* have no perceived consequences for *B,* whereas in the other (*C*) those actions do have perceived consequences for *B*. *C** will then be the *solitary* counterpart of *C,* and *C* will be the *social* counterpart of *C**. If *A* forms different desires in *C** from those *A* forms in *C,* the set of desires present in *C** will be *A*'s *solitary* desires (relative to the counterparts *C* and *C**). Given these preliminary specifications:

> *A* acts psychologically altruistically with respect to *B* in *C* just in case
>
> (1) *A* acts on the basis of a desire that is different from the desire that would have moved *A* to action in *C*,* the solitary counterpart of *C*.
> (2) The desire that moves *A* to action in *C* is more closely aligned with the wants *A* attributes to *B* in *C* than the desire that would have moved *A* to action in *C**.
> (3) The desire that moves *A* to action in *C* results from *A*'s perception of *B*'s wants in *C*.
> (4) The desire that moves *A* to action in *C* is not caused by *A*'s expectation that the action resulting from it would promote *A*'s solitary desires (with respect to *C* and *C**).

Condition 1 tells us that *A* modifies his or her desires from the way they would otherwise have been, when there is an impact—more accurately, when there is a perceived impact[9]—on the wants of *B*. Condition 2 adds the idea that the desire, and the behavior it directs, is more in harmony with the wants attributed to *B* than it would have been if *B* were unaffected by what was done. (It is possible to modify your desires in response to the perceived wishes of another, but to do so in a way that *diverges* from their perceived wants—that is spite.) Condition 3 explains that the increased harmony comes about because of the perception of *B*'s wants; it is not, say, some caprice on *A*'s part that a different desire comes into play here. Finally, condition 4 denies that the modification is to be un-

9. I shall consider cases in which agents have mistaken beliefs later. For the time being, I suppose that the parties get things at least roughly right.

derstood in terms of *A*'s attempt to promote some desire that would have been present in situations where there was no thought of helping or hurting *B;* this distinguishes *A* from the food sharer who hopes for returns on future occasions when *B* is in the position of disposing of the goods. Condition 4 requires that genuine psychological altruists be different from Machiavellian calculators who aim to satisfy the wants they would have in solitary situations (I shall sometimes refer to condition 4 as the "anti-Machiavelli" condition).

Given this account of psychological altruism, it is now possible to characterize *behavioral* altruism more carefully. Behavioral altruists are people who look like psychological altruists. That is, they perform the actions people with psychologically altruistic desires would have been led to perform. In ascribing behavioral altruism, however, we do not suppose any particular psychological explanation of the actions. Perhaps they are indeed the products of psychologically altruistic desires, or perhaps the actions are produced by quite different desires having nothing to do with the satisfaction of the beneficiary—a desire for status, or for feeling oneself in accordance with some socially approved pattern of conduct, or even a self-interested calculation. (We shall explore some possibilities of behavioral altruism later; see §§7, 11.)

The stylized food example allows the introduction of an obvious concept, one that will be important in future discussions, and that further articulates the account of altruistic desires. The altruistic modification of solitary desires can be more or less *intense.* I have spoken—somewhat vaguely—of the altruist as aligning his or her wants with those attributed to the beneficiary.[10] That alignment is often a matter of degree, for example, when there is a continuum of possibilities intermediate between complete egoism (retaining one's solitary desires in the social counterpart) and complete subordination of one's solitary wishes to those one perceives the other to have (where one comes to want exactly what one perceives the other as desiring). In sharing food, this is easily

10. For a more precise and formal discussion of many aspects of altruism, see my essays "The Evolution of Human Altruism," *Journal of Philosophy* 90 (1993): 497–516, and "Varieties of Altruism," *Economics and Philosophy* (2010): 121–148. As I shall note at various places, there are several aspects of the account of altruism provided in this chapter that can be treated mathematically, and these articles make a start on that.

expressed in terms of the mode of division: egoists give nothing, self-abnegating altruists give everything, and in between lie a host of intermediate altruists. One obvious style of altruism is *golden-rule altruism,* distinguished by its equal weighing of the solitary desires and those attributed to the beneficiary.

Inspired by the food example, we can undertake a simple way of representing the intensity of psychological altruism, one that will be useful in some (but by no means in all) instances. Suppose that people's desires can be represented by (real) numbers that correspond to how much they value a given outcome. If one result, eating all the food, say, is worth 10 to me, and another, eating half the food, is worth 7, then I prefer eating everything to eating half, but I also prefer an assured outcome in which I receive half to the state of being awarded all or nothing dependent on the flip of a fair coin. (For, in the latter case, my expected return is measured by 5—half of 10 plus half of 0—which is less than 7.)

When you are in the picture, I also take into account the values you attribute to various outcomes. My social desire could be represented as a weighted average of the values represented in my solitary desires and those I take to measure your solitary desires. Thus, the numbers assigned in my social desires would be given by the simple equation:

$$v_{Soc} = w_{Ego} v_{Sol} + w_{Alt} v_{Ben}$$

where v_{Soc} measures my social desires, v_{Sol} my solitary desires, v_{Ben} the measurements of desire I attribute to the beneficiary (you), and w_{Ego} and w_{Alt} the weights given to my solitary desires and my attributions of desire values to you (so that $w_{Ego} + w_{Alt} = 1$). The intensity of my altruism is represented by the size of w_{Alt}—and hence inversely by the size of w_{Ego}; if $w_{Ego} = 1$ ($w_{Alt} = 0$), then I am, at least with respect to you on this occasion, a psychological egoist; if $w_{Alt} = 1$, then I am a self-abnegating altruist; if $w_{Alt} = 0.5$ ($= w_{Ego}$), then I am a golden-rule altruist.

We should not assume that all types of altruistic alignment with the wishes of others can be conceived in this very simple way. Cases of sharing show that a simple approach sometimes works, and the simple expression of social wants as weighted averages will be useful in explaining and illustrating some of the ideas of later sections.

§4. The Varieties of Altruistic Reactions

As already recognized, altruism is not always about the modification of desire, though we are often reasonably suspicious about alleged examples of altruism that do not change desires in ways leading to action: it is not enough simply to "feel another's pain." We can be moved to share the hopes of others, to modify our own long-term intentions and aspirations to accommodate what we see them as striving for, and, most important, we can feel different emotions because of our awareness either of what they feel or of the situations in which they find themselves. For some kinds of psychological states, hopes and long-term intentions, for example, accounts of altruistic versions of these states can be generated straightforwardly in parallel to the treatment of the previous section. Emotions, however, deserve special consideration, both because they are frequently components of the psychological attitudes with which we shall be concerned, and because they involve types of reactions more broadly shared among animals than the psychological states on which I have so far concentrated.

Altruistic emotional responses to others *might* be—and probably often are—mediated by perception and cognition. We see that another person is suffering—or jubilant, for altruistic emotions are not always dark—and our own emotional state changes to align itself more closely with that attributed to the other. Or, in a different mode of altruistic response, we understand the situation in which another person is placed, and our emotional state changes to take on, to some extent, the feeling(s) we would have if placed in that state.[11] When people, or other animals, have dispositions to modify their emotional states in light of their understanding of the feelings or the predicaments of others, we can treat emotional altruism just as §3 analyzed altruistic desire. The emotional altruist feels one thing in the solitary counterpart and feels differently in the social counterpart; the emotion in the social counterpart is more closely aligned with that attributed to the other (or more closely aligned

11. The distinction between these two modes of altruistic emotional response—forms of "sympathy"—was already clearly recognized by Adam Smith in *Theory of Moral Sentiments,* Knud Haakonssen, ed. (New York: Cambridge University Press, 2002).

with the emotion the altruist supposes he or she would feel if placed in the other's shoes), and the alignment comes about because of the recognition of the other's feelings (or of the other's situation); finally, it is not caused by any background solitary emotion or solitary desire. Now, whereas in the understanding of altruistic desire this last condition responds to a genuine worry—for we readily think people can form ostensibly other-directed desires on the basis of selfish calculations (I can want to share with you because I think it will be good for me in the long run)—the anti-Machiavelli condition seems odd and gratuitous in the emotional case. It is natural to think, and it may even be true, that self-directed psychological states simply have no power to generate *emotions* toward others, that our emotional life is not under that sort of control. Emotional responses, one may suppose, are caused by processes more direct and automatic than the perceptions and cognitions figuring in my analyses. Consequently, an account of emotional altruism parallel to the analysis of altruistic desire will be at least incomplete, and perhaps even radically misguided.

This is a serious challenge. To meet it, we shall have to consider, if only briefly, the character of emotions. Without taking sides in unresolved controversies, I shall argue that some kinds of emotional response can be understood along the lines just sketched, while others cannot. An account of more basic altruistic emotional reactions, or "affective states," as I shall call them, provides a valuable supplement to the approach to psychological altruism begun in the previous section.

Emotions involve changes in our physiology, and some students of emotion have identified the emotion with the alteration in physiological state. Others propose that there are important distinctions among emotions that cannot be recognized without supposing those who feel the emotions to have particular beliefs, desires, and intentions: specific forms of awareness are required for guilt and shame, for resentment and indignation, and for certain kinds of contentment and anger. A natural way of responding to the findings of neuroscientists, psychologists, and anthropologists is to suppose that many emotions are complex entities, perhaps processes in which particular types of physiological conditions are accompanied by special kinds of cognitive and volitional states. When someone resents the insensitive remarks made by another, he or she undergoes

a physiological response connected to judgments about what has been said and desires about what will happen next. The causal details of these connections are matters of speculation, but, even in advance of knowing them, we can reject an approach to emotions that would leave out either the physiological or the cognitive/volitional features.[12]

Yet there may be emotional states, felt by nonhuman animals and by human infants perhaps, for which the cognitive/volitional component is negligible, even entirely absent. With respect to our own species, it has been argued that there are a number of *basic* emotions, found in all human societies and typically giving rise to the same facial expressions.[13] Although a widespread aspect of human psychology or behavior is often taken as evidence of some biological (typically genetic) basis that generates the common feature across all environments, it is worth treading carefully here. For, trivially, there will be some environments in which members of our species will not develop so as to exhibit the typical reaction—neural and psychological development can be disrupted in many different ways. The interesting questions are whether there are subtle properties of potential human social environments capable of prompting something different (so that the widespread finding of the common feature depends on the absence of those subtle properties in the human societies studied), and whether, if so, the potential environments are in some way pathological. These questions are not settled, but, for the sake of the present discussion, I shall allow that human beings who develop in environments, physical and social, that do not involve damaging disruptions of development, will all share dispositions to *basic affective reactions*—that is, they will have capacities for basic affective states, like disgust and anger and fear, and they will exhibit similar facial expressions characteristic of the individual affective states.[14]

12. Here I am much influenced by the thoughtful and ecumenical approach adopted by Jenefer Robinson in the first three chapters of *Deeper than Reason* (New York: Oxford University Press, 2005).

13. See Paul Ekman and Richard Davidson, eds., *Nature of Emotion* (New York: Oxford University Press, 1994).

14. Since the role of social environments is central to my approach to our altruistic tendencies and the character of the ethical project, my position would be strengthened if this concession proves false.

This concession does not entail any of a number of conclusions sometimes drawn from it. First, there is no implication that the affective reactions in people who belong to different societies will be generated by the same things, events, and states of the world: virtually all people feel disgust, but different groups find very different things disgusting. Second, to allow basic affective reactions does not retract the earlier judgment that many emotions have cognitive/volitional components: every emotion might involve some basic affective state, but a large diversity of emotions might be distinguished by the cognitions and volitions connected to that state. Third, and most important for present purposes, we should resist the thought that "because they are biological," affective reactions are based on some mechanism more "immediate," more "primitive" than human cognition. It is easy to muddle together two different senses in which a capacity may be "biological," one in which its development occurs across all (nonpathological) environments, and one in which the ways in which it is activated bypass our beliefs and wishes. Conceding the "biological" status of affective reactions in the first sense does not commit us to supposing them "biological" in the second.

We can now address the critical issue concerning altruistic emotions. Even though it is not *required* that there be affective reactions that do not depend on the causally prior recognition of the feelings or the predicaments of others (on *beliefs* that they are suffering, for example), it is *possible* that there should be mechanisms prompting particular affective responses, mechanisms not mediated by prior cognition. It is well-known that infants in the same hospital nursery react to the crying of others: an initial solo can set off an entire chorus.[15] Supposing the unhappiness of the one spreads to the many because each believes that someone around is unhappy strains credulity. A more sober account would view the process as a kind of contagion, effected by a species-typical neural mechanism, that transfers the misery of one baby to those around. Recent studies of the activity of so-called mirror neurons (primarily investigated in macaques) may offer clues about the potential mechanism. Perceptions, or even sensations, can cause an animal to activate the same neurons as

15. See Martin Hoffman, *Empathy and Moral Development* (New York: Cambridge University Press, 2000).

those giving rise to the behavior perceived or causing the sensations: *A*'s observation of *B*'s facial expression produces neuronal firings that tense the pertinent muscles and that result in *A*'s imitation of *B;* perhaps the sound of another baby crying induces a pattern of neural activity that mirrors that in the source of the crying and thus causes the originally contented baby to cry.[16] Mechanisms of this sort require a different approach to altruistic emotions.

Once we have the challenge clearly in view, however, it is not hard to see how to liberate the account of altruism begun in §3 from its dependence on cognition. The task is to provide a definition of "*A* feels an altruistic emotion in response to *B* in *C*." As before, we shall suppose the notions of solitary and social counterparts. The conditions are as follows:

1. *A* feels an emotion different from the emotion *A* would have felt in *C*,* the solitary counterpart of *C*.
2. The emotion *A* feels is more closely aligned with the emotion *A* attributes to *B* in *C* than the emotion *A* would have felt in *C*;* or it is more closely aligned with the emotion *A* would have felt had *A* been in *B*'s position in *C;* or it is more closely aligned with the emotion *B* actually feels in *C*.
3. If the emotion *A* feels is more closely aligned with that attributed to *B* or if the emotion *A* feels is more closely aligned with the emotion *A* would have felt in *B*'s position in *C,* then the emotion *A* feels in *C* results from *A*'s perception of *B*'s situation in *C;* if no recognition of *B*'s state plays a causal role and if *A*'s emotion aligns itself with that felt by *B,* then the emotion felt by *A* is caused by the operation of some automatic neural mechanism, a mechanism triggered by *A*'s observation of *B* (the activation of mirror neurons might be one such mechanism).
4. The emotion felt by *A* in *C* is not caused by *A*'s expectation that feeling this emotion would promote *A*'s solitary desires (with respect to *C* and *C**).

16. William Damon, *The Moral Child* (New York: Free Press, 1998); and Hoffman, *Empathy and Moral Development.*

This account introduces clauses into the second and third conditions in order to allow the possibility of altruistic emotions produced in ways that bypass cognition. Although the fourth condition is retained, it is highly plausible that Machiavellian manipulation of our emotional lives is beyond our powers, and, if that is indeed so, this requirement is redundant.

The analysis just given preserves a fundamental feature of my original characterization of psychological altruism (§3): altruists have a particular type of relational structure in their psychological lives—when others are around, the altruist's desires, hopes, intentions, and emotions are different from what they would otherwise have been, closer in some way to those of the others, and the difference is produced by some sort of response to those others, not by something enclosed within the self (calculations of future benefit, for example). What the more complex approach to altruistic emotions adds is the possibility that the generation of the response might involve some precognitive mechanism.

It is easy to overinterpret this last point. One might suppose that affective states are always generated by some mechanism that does not involve cognition—but, not only do I see no basis for holding so sweeping a generalization, but it also seems belied by the fact that affective reactions are often founded in complex and explicit understanding (when I see pictures of Jewish refugee children being greeted at English ports by policemen and willing foster parents, I feel a complex mixture of emotions, surely involving affective states, but these states are clearly dependent on my conscious understanding of what the photographs display). The causal relations among affective and cognitive states may be quite various, and, while we await definitive accounts of them, it is well to suspend judgment and to be open to many possibilities.

Nor should we suppose that noncognitive mechanisms are inevitably involved in whatever altruistic responses occur in nonhuman animals. Although questions about the extent of animal abilities to recognize the wishes and thoughts of their conspecifics are much debated, there is no reason to take an advance stand on these issues.[17] I shall later defend the

17. For defenses of opposing views, see de Waal, *Primates and Philosophers;* and Derek C. Penn and Daniel J. Povinelli, "On the Lack of Evidence that Non-human Animals Pos-

thesis that some of our evolutionary cousins have altruistic desires (in the sense of §3; see §7) and that similar capacities were shared by our hominid ancestors.

§5. Some Dimensions of Altruism

One further aspect of psychological altruism needs to be emphasized before we have all the tools required for probing the hominid preethical state. On the account of the last sections, there are many varieties of altruism. Or, to use a suggestive metaphor, altruism is a multidimensional notion. For animals capable of psychological altruism, each individual occupies a particular place in a multidimensional space where brute (non-Machiavellian) egoism is represented by a single plane, and the various forms of altruism range over the entire rest of the space.[18]

An animal's *altruism profile* (where he or she is located in altruism space) is determined by five factors: the *intensity* of the animal's responses to others, the *range* of those to whom the animal is prepared to make an altruistic response, the *scope* of contexts in which the animal is disposed to respond, the animal's *discernment* in appreciating the consequences for others, and the animal's *empathetic skill* in identifying the desires others have or the predicaments in which they find themselves. Non-Machiavellian egoists never respond to anyone else in any context: for the dimensions of intensity, range, and scope they score 0, 0, and 0; their discernment and empathetic skill can be as you please, for these are never called into play.

Altruists are not like that. They modify their desires and emotions to align them with the perceived desires and (perceived or actual) emotions of at least some others in at least some contexts. As §3 already proposed, their responses may be more or less intense. With respect to altruistic desires, an altruist may give more or less weight to the perceived desire of the beneficiary. My treatment of the stylized example in terms of weighted averaging provides a clear paradigm for intensity—the intensity

---

sess Anything Remotely Resembling a 'Theory of Mind,'" *Philosophical Transactions of the Royal Society B* 362 (2007): 731–44.

18. For more details about this spatial metaphor, see Kitcher, "Varieties of Altruism."

of altruism is represented by how much of the food you are willing to relinquish. If

$$v_{Soc} = w_{Ego} v_{Sol} + w_{Alt} v_{Ben}$$

egoists set w_{Ego} at 1 and w_{Alt} at 0. People for whom $w_{Ego} = 1 - \varepsilon$, where ε is tiny, are altruists in a very modest sense: they will act to advance the wishes of others only when the perceived benefits to others are enormous compared to the forfeits for themselves—they may suffer the scratching of their finger in order to avoid the destruction of the world, but refuse larger sacrifices. People for whom $w_{Alt} = 1$, by contrast, are completely self-abnegating. They abandon their own solitary desires entirely, taking on the wishes they attribute to the beneficiary. In between, we find golden-rule altruists, for whom $w_{Alt} = 1/2$, who treat the perceived wishes of the other exactly as they do their own solitary desires.

Even when averaging is not appropriate for representing altruistic desires, there will often be a comparable notion of the degree to which one has accommodated the perceived wishes of the other. Moreover, with respect to altruistic emotions there is surely a similar concept. Notoriously, we can be relatively unsympathetic, even with those who are dearest to us, when we are preoccupied or distracted. At other times, we enter fully into the feelings of friends and loved ones, even of strangers. It is not obvious how to delineate the notion of intensity in the emotional case as precisely as the food-sharing example allows, but the varying intensity of altruism in emotional responses is uncontroversial. Notice, however, that it should not be confused with the intensity of *emotion:* intensity depends on the degree of *alignment* with the other's feelings (or with the feeling one would have had in the other's situation), not with the force of what one feels.

Most altruists, indeed probably all, lack a fixed intensity of response, applying with respect to all potential beneficiaries and all contexts. There are many people to whom we would rarely make an altruistic response: these people effectively fall outside the range of our altruism. Even with respect to those to whom we are disposed to respond, there are many contexts in which we do not take their perceived wishes or their feelings into consideration (or into our own minds). For many, perhaps,

we are prepared to offer limited forms of aid and support; for a few, we are willing to sacrifice everything. Often our altruistic responses to some are colored by indifference to others: parents who make sacrifices to help their children obtain things the children passionately want frequently do not take into account the wishes of other children (or the altruistic desires of the parents of the other children).

Someone's altruism profile typically shows a relatively small number of people to whom the focal individual responds, frequently with significant intensity, across a wide set of contexts. The beneficiaries lie at the center of the range of altruism for the focal individual, and the scope for these beneficiaries is wide. As we consider other potential beneficiaries more distant from the center, the scope narrows (there are fewer contexts in which the more peripheral people elicit an altruistic reaction) and the intensity falls off, until we encounter people to whom the focal individual makes no altruistic response at all. Henceforth, I shall conceive of the range of *A*'s altruism in terms of the metaphor of center and periphery: the center is the select set of potential beneficiaries for whom *A*'s response is relatively intense across a relatively wide scope of contexts; at the periphery, the intensity of the response and the scope of contexts narrow and vanish.

Someone's character as an altruist is not fixed simply by the factors so far considered—intensity, range, and scope—because there are also significant cognitive dimensions to altruism. *A* may make no response in a particular context through failure to understand the consequences for *B;* perhaps *A* does not differentiate the social from the solitary counterpart. Often this is an excusable feature of our fallibility, for the impact on the lives of others may be subtle; we may just not see that following some habitual practice—buying at the most attractive price, or investing in promising stocks—has deleterious consequences for people about whose welfare we care. Evidently, however, acuity with respect to consequences comes in grades, and we admire those who appreciate the intricate ways in which others can be affected, while blaming those who "ought to have seen" the damage they cause.

Similarly, there are degrees to which people are good at gauging the desires of others. Almost everyone is familiar with the well-intentioned person who tries to advance the projects of an intended beneficiary but

who is hopelessly misguided about what the beneficiary wants: almost everyone has had a friend or relative who persists in giving presents no longer appropriate for the recipient's age or conditions of life. It would be hard, I think, to declare that people who attribute the wrong desires to their beneficiaries, or who overlook consequences for those whom they intend to benefit, are not acting altruistically when they carry out their variously misguided plans—their intentions are, after all, directed toward doing good for others—but their altruism needs to be differentiated from that of their more acute fellows. Hence I add two cognitive dimensions, one representing *A*'s skill in understanding the nature of a social counterpart to a solitary context, and one assessing *A*'s ability to empathize with *B,* to ascribe desires *B* actually possesses.

A simple reaction to the prospect of human egoism is to propose that people living in community with one another—or even all people—should be altruistic; some even take the second commandment of the New Testament to constitute a complete ethical system. Recognizing the dimensions of altruism undermines that thought. There is no *single* way to be an altruist, and, consequently, the commendation of altruism must be given more specific content. What kind of altruist should we urge someone to be? Moreover, is it right to suppose that the best state of the community (or the entire species) is achieved by having each member (each person) manifest the same altruism profile? You might think the questions have straightforward answers. Along the cognitive dimensions, accuracy is always preferable: ideally people should be aware of the potential impact for others and should understand what others want. For issues of intensity, range, and scope, we ought to aim at golden-rule altruism with respect to all people across all contexts.

The demand for accuracy on the cognitive dimensions is more plausible but still not uncontroversial. Debate about the second part of the proposal arises in obvious ways. It might be valuable for people to develop strong ties with some others—the range of human altruism should have a definite center; from Freud's worries about the "thinning out" of our libido in the development of civilization to familiar philosophical examples about parents who wonder whether they should save the drowning child who is closer, when their own drowning child is farther out and harder to rescue, a spectrum of troublesome cases arouses suspi-

cion about completely impartial altruism.[19] Moreover, in a world with finite resources, the desires of others often conflict. If *A* accurately perceives that both B_1 and B_2 want some indivisible good, it should not be automatic that *A*'s desire should be formed by treating B_1 and B_2 symmetrically. (We may, for example, want *A* to respond to aspects of the history of the situation, including what B_1 and B_2 have previously done.) None of this is to deny that there may be a level at which we want altruism profiles to respond impartially to others, but merely to insist that the impartiality we want cannot be adequately captured as golden-rule altruism toward all people in all contexts.

Further complexities of the notion of psychological altruism will occupy us later. For the present, however, we have enough to begin charting the history of our ethical practices, by understanding how the most basic forms of psychological altruism could have evolved, and how they formed an important part of the social environment in which the ethical project began.

§6. Maternal Concern

Before our human ancestors invented ethics, they had a capacity for psychological altruism. This thesis might be disputed in any of several ways, but the one of immediate concern recapitulates the skepticism about altruism mentioned earlier (§3). Armed with the elements of an account of psychological altruism, the first task is to decide if any such capacity exists, and if it could plausibly be attributed to contemporary human beings, our hominid ancestors, and our evolutionary cousins. Let us begin with the most straightforward case.

Behavior directed toward the survival of young is quite widespread in the animal kingdom, found, for example, among birds as well as mammals. With respect to some types of animals, the hypothesis that this behavior is directed by altruistic desires appears extravagant, for it pre-

19. Sigmund Freud, *Civilization and Its Discontents* (New York: Norton, 1961); Bernard Williams, "Personhood, Character and Morality" in *Moral Luck* (Cambridge, UK: Cambridge University Press, 1981); Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," in *Facts, Values, and Norms* (Cambridge, UK: University Press, 2003).

supposes the propriety of attributing wants and intentions apparently beyond the cognitive capacities of the pertinent organisms. Nevertheless, we might view the animals as driven by altruistic emotions (or primitive versions of them), generated through the operation of automatic neural mechanisms. Among primates, however, particularly those closest to our own species, our evolutionary cousins the great apes, there is considerable evidence for the ability to have desires and to recognize the desires of others.[20] For the sake of concreteness, we can think of psychologically altruistic dispositions to care for the young as emerging in apelike ancestors of *Homo sapiens,* but it is eminently possible that they evolved much further back in our primate (or even mammalian) past.

Even those who share the orthodox primatological views about the cognitive sophistication of our evolutionary cousins may be skeptical of any hypothesis that parental care is sometimes directed by altruistic desires, in the sense I have explicated in §3. They may wonder, for example, whether any dispositions of this kind could evolve under Darwinian natural selection, or whether the apparently altruistic behavior is really the product of some quite different mechanism. Perhaps the animals are really calculating how to achieve future benefits, violating condition 4 of my account, the anti-Machiavelli condition. Many primatologists take the social organization of primate life to reveal "Machiavellian intelligence," and evolutionary psychologists often propose that increased cognitive powers in hominids reflect the need to manipulate others and to avoid being manipulated oneself.[21] Or perhaps the plausible candi-

20. There are many excellent sources for attributing complex cognitive states to nonhuman primates. See, for example, Dorothy Cheney and Robert Seyfarth, *How Monkeys See the World* (Chicago: University of Chicago Press, 1990), esp. chaps. 3 and 8; Jane Goodall, *The Chimpanzees of Gombe* (Cambridge, MA: Harvard University Press, 1986); C. Bachmann and H. Kummer, "Male Assessment of Female Choice in Hamadryas Baboons," *Behavioral Ecology and Sociobiology* 6 (1980): 315–21; R. Byrne and A. Whiten, eds., *Machiavellian Intelligence* (Oxford, UK: Oxford University Press, 1988), particularly the essay by Nicholas Humphrey ("The Social Function of Intellect," 13–21).

21. Many, though not all, of the essays in Byrne and Whiten, *Machiavellian Intelligence* (see n. 20), adopt this perspective. For a more pronounced articulation of the theme that intelligence is a tool for calculating egoists, see James Barkow, Leda Cosmides, and John Tooby, eds., *The Adapted Mind* (New York: Oxford University Press, 1992). In "The Social Function of Intellect," Nicholas Humphrey offers a broader vision (see esp. p. 23).

dates for altruistic responses to the young are affective and immediate. That would allow for altruistic emotions, even emotions that direct behavior, but not necessarily for altruistic desires. To address this latter concern, I shall begin with an example that involves serious cognition and planning.

Primates roaming on the savannah sometimes encounter carcasses that could serve as food. Imagine a female finding a carcass in the absence of her young. Instead of devouring it on the spot, she quickly summons her young. It is difficult to think of behavior of this sort as an action driven by instincts or emotions. Apparently, the mother has to recognize this as food she can share, and to prefer sharing to devouring it entire. Perceiving the possibilities for her young, she forms a different desire from the one she would have formed had they been out of range or fully mature and dispersed. That desire underlies her efforts to summon them to the scene before the food spoils or is taken by another animal. On the face of it, this is an example of altruistic desires in the sense of §3.

One line of concern about attributing altruistic desires is that capacities for such wishes could not have evolved and been maintained under natural selection. In settling this worry, we can use the tools supplied to solve the problem of biological altruism. Suppose that food has decreasing marginal value (in terms of promoting reproductive success), so that, although eating a whole carcass has a higher effect, on fitness it is considerably less than double the effect of eating just half. Assume that the mother has a disposition to golden-rule altruism (or some approximation of it) with respect to her offspring, and that there is just one of her young in the vicinity. Then it is not hard to show that this disposition can be favored by kin selection.[22]

The more difficult challenge asks whether all the conditions for psychological altruism have been met. Perhaps the adjustment of desires to accommodate the perceived needs of young is based upon "Machiavellian" calculations. What form might these supposedly self-directed processes take? Begin with a style of skeptical argument rarely made explicit, but one underlying the conviction that references to psychological altruism are exercises in sentimental self-deception. According to

22. For details, see Kitcher, "Evolution of Human Altruism."

this line of thought, the benefits to offspring, favoring the evolutionary success of altruism, undermine its genuineness. In the described scenario, however, the mother must do something psychologically sophisticated—she has to recognize this as an occasion for seeking out her young—rather than simply exhibit some instinctive reaction. What, then, is the alternative cognitive account that replaces the disposition to adjust preferences with Machiavellian calculation? It strains credulity to suppose mothers recognize the evolutionary advantages of sharing: only a few very select primates *could* calculate the genetic gains and losses (and those who do make their judgments in this way are, to say the least, misguided). So if she calculates it will have to proceed via proxies, through the attempt to attain selfish goals correlated with increases in reproductive success. What could those be?

The most plausible answer is that maternal care proceeds from expectations of future reciprocity—the child is expected to grow into a future ally, maybe eventually a caregiver. Here, the consequences of the present action would be represented in terms we can imagine being within the mother's conceptual repertoire, but we are supposing animal abilities to abstract from present conditions and to envisage a very different future, to overlook the weak juvenile and see a future strong ally. Even if we allow such amazing foresight, problems remain. If dispositions to share with young evolve under natural selection because of inclusive fitness considerations, then the expectations of future aid ought not to be an accurate guide to the kinds of behavior selection would favor—the alleged proxies do not match up well with the variables (the gene frequencies) that are the "ultimate currency" of evolution. From the standpoint of inclusive fitness, mothers should provide some aid when there is very little chance of reciprocity in the future (simply because, even without reciprocation, helping offspring is a good way to spread the genes), and they should provide extra aid to offspring who can be expected to reciprocate. If the hypothetical calculation is to give values that correlate with inclusive fitness, the perceived gains from reciprocity have to be inflated. Why should mothers think their care will be remembered, or, if recalled, it will trigger a disposition to repay? If sharing is based on the expectation of returns, the young seem bad targets. Other, more mature, members of the group would appear to be better prospects for future aid.

The best version of skepticism invokes psychological variables correlating closely with the well-being of the young, and thus presumably with the spread of the pertinent alleles. Determined "Machiavellians" may concede that the scenario I described—in which mothers bring young to share food—involves cognitive abilities, but they may view the calculations that occur as directed toward benefits arising from simpler, more instinctive, reactions. Start, then, with maternal responses to distress. Here, it might be alleged, mothers promote their own ease by preventing wails, facial expressions, and upsetting bodily gestures; or, more positively, mothers find psychological pleasure in observing smiles or hearing happy gurgles. This, it is conceded, is a form of *emotional* altruism. Hence, on occasions where offspring are present, maternal behavior (hugging, caressing, giving food) is directed by the desire either to avoid an unpleasant state ("the pang," generated from emotional responses to unhappy young) or to attain a pleasant one ("the glow," similarly generated from emotional reactions to contented young).[23] When the young are not directly present, however, but available to be brought to the carcass, the mother recognizes the possibility of attaining the glow (by bringing them to the scene) or the dangers of experiencing the pang (if she devours the whole carcass and then encounters hungry offspring). So she calculates that her own selfish desires can better be satisfied by sharing. Because the anti-Machiavelli condition is violated, she does not count as a psychological altruist.

At least two problems confront this skeptical response. The first, and more obvious, is the highly implausible style of cognition it attributes to the mother at the scene of the carcass. She is supposed to be capable of representing to herself not only her absent offspring and their need for food (as on the interpretation of her as a psychological altruist), but also the ways potential actions will bring about glows and pangs—she has to have such thoughts as "If I find the young and share, I shall enjoy the glow" or "If I devour all the food, I shall suffer a pang when I meet the

23. In *Unto Others* (Cambridge, MA: Harvard University Press, 1998), Elliott Sober and David Sloan Wilson rightly regard this kind of skeptical response as the most important challenge to the existence of psychological altruism. I think their way of dealing with it is unnecessarily complex, and offer a simpler treatment. Nonetheless, we are in agreement that the challenge can be met.

offspring." Even the most liberal cognitive ethologist is likely to wonder if thoughts like these are within the repertoire of our primate relatives. Moreover, to deliver the appropriate behavior, the envisaged glow or pang has to be sufficiently vivid to override the present desire for the available meat. Only anti-altruist prejudice could inspire the idea that these hypothetical calculations plausibly reconstruct the animals' psychological lives.

The story already presupposes one type of altruistic tendency: mothers feel altruistic emotions. That was allowed in describing the situations from which the skeptical response sets out, the distress felt in the presence of howling infants, pleasure in smiles and gurgles. By the skeptic's own lights, altruistic emotional responses (in the sense of §4) underlie the Machiavellian calculation. Curiously, the skeptical complaint assumes that these emotional responses engender complicated cognitive and volitional states (beliefs and desires about glows and pangs) but do not issue in much simpler desires. The mother's emotional response to her needy young produces no desire to feed them, but a longing for glows or a fear of pangs. Invoking complex Machiavellian calculation and ignoring the far simpler psychological route leading from emotion to simple desire again looks like an egoist prejudice, not a serious rival hypothesis.

These points can be developed further by temporarily leaving our evolutionary past and focusing on apparent altruism among human parents. Imagine a mother whose child has some serious need, a need difficult to satisfy—the child must be rescued; the mother has to engage in an intricate and risky procedure to have any chance of saving the child's life. Enough determined mothers pursue similar causes with unusual energy and persistence, and for them hypotheses about future reciprocation, respect from third parties, or enhanced social status would be jokes in extremely poor taste. The most difficult form of the skeptical hypothesis proposes that these mothers are driven by internal mechanisms—particularly by desires to avoid the pang. We find it natural to suppose that they "couldn't live with themselves" unless they did everything possible for the child (interestingly, in the human case, we tend to recognize the supposed psychological states, the glows and pangs, as intertwined with matters of conscience, a point that will be important later).

Hence, the skeptic proposes, mothers do the impressive things they do because they want to avoid a future of terrible self-reproach and self-torment.

At least two things cast doubt on the skeptical hypothesis. First, the fact that the mother envisages the future of self-reproach testifies to the motivating power of her recognition of the child's wishes (or, in this instance, more likely the child's *interests*—see §21). It is often preposterous to suppose a mother will reproach herself because she is concerned with attitudes in her society—frequently, those around her would praise her for doing far less than she does, constantly reassure her that she has done more than anyone could possibly expect, and so forth. The drive to pursue every possible avenue comes from within, and it could not be abated by any amount of well-intentioned commendation and comfort. If she fails, the mother will suffer, no matter how much she has done and no matter what others say, and the suffering will stem from her deep desire that the child survive and flourish. So, at least, we might initially believe. On the skeptical hypothesis, however, that desire must be denied. Instead, the mother must be viewed as being able to feel altruistic emotions in response to her child. This ability, and the emotions to which it gives rise, does *not* express itself in a desire for the child's well-being. Instead, the ability leads her to fear a particular type of future state, and the fear replaces the denied desire as the driver of her conduct. We have no grounds for accepting this speculative psychology.

A final—fanciful—way to underscore the point: Our world hardly abounds with clever spirits, willing to offer bargains. Yet the mother might have a particular disposition to react to temptations. Imagine that she were visited by a Mephistophelean figure with a straightforward proposal: "I can give you a pill to ensure you will not feel any guilt should things go badly for your child. The pill will wipe away both the pangs of conscience—you will reflect on your efforts and feel you did your best—and any memory of this conversation and the decision to accept the pill. The downside is truly tiny. The probability of your saving your child if you don't take the pill is p; the probability if you do take the pill is $p - \varepsilon$ (where ε is really infinitesimal). Surely the reasonable thing is to accept?" With respect to many actual mothers, we have no doubt about how they would respond—by telling Mephisto to get lost. They view their future

psychological comfort as trivial compared with the value of saving the child—any diminution of the probability of success is a loss for which future amnesia cannot compensate. Their assessment of relative value expresses just the desire for the child's well-being the skeptic attempts to deny.

Psychological altruism is real, it is exemplified in maternal concern, and it originally evolved through the most fundamental type of kin selection. Because it is hard to envisage how psychological altruism could take hold without directing maternal care—no other social bond is as pervasive in our evolutionary past, no other recurrent situation is as relevant to reproductive success—it is the most basic and primitive type of altruism.

§7. Broader Forms of Altruism?

How far does psychological altruism extend? Is it merely something mothers (or parents) direct toward their young?

For a first example, we can turn to the inverse of the relationship just examined, to occasions on which offspring help their parents. In her study of the chimpanzees of Gombe, Jane Goodall relates a moving story about the behavior of an adult female, Little Bee, who tended to her partially paralyzed mother, Madam Bee.[24] On several occasions, Little Bee and her mother lagged behind the rest of the troop, often arriving at the nesting site hours later than the others. Mother and daughter took frequent rests, and, when food was needed, Little Bee climbed trees, collecting fruit to share with Madam Bee. Apparently, Little Bee adjusted her preferences to accommodate the perceived needs of her mother, and by doing so she exposed herself to risks she might otherwise have avoided. Reading Goodall's account, it seems clear that the first three of my conditions for psychological altruism are satisfied. The crucial requirement, where skepticism so often arises, is the anti-Machiavelli condition.

Was Little Bee's adjustment of her preferences based on calculating some narrow advantage for herself? It is hard to think what it might be.

24. Goodall, *Chimpanzees of Gombe*, 357, 386.

There was no realistic possibility of her pronounced efforts on behalf of her mother being reciprocated by some future benefits conferred by Madam Bee. Nor could she obtain extra status among the members of her troop, who were in no position to witness her actions—indeed, because her time for interacting with others in the group was so drastically curtailed, her chances for cooperative interactions with them were diminished. If her behavior resulted from calculation, aimed at advancing her own solitary wants, the only possible conclusion is that she *mis*calculated, but the miscalculations would have been so gross as to be quite at odds with her demonstrated social intelligence. Far more plausible is the hypothesis that Little Bee was what she seemed to be—a psychological altruist.

Similarly for a young male chimpanzee, observed by Frans de Waal:[25] Early one morning, de Waal watched two members of the Arnhem chimpanzee colony enter the outdoor enclosure: Krom, a somewhat retarded mature female, and Jakie, a healthy young male. It had rained overnight, and rain had collected in one of the tires hanging from a horizontal pole attached to the climbing frame. Krom wanted to free that tire, but, unfortunately, it was the innermost of five, and her efforts at removing all five tires at once proved futile. After she sat down in a corner of the enclosure, Jakie approached the frame. Intelligently, he removed the tires one at a time, carefully carried the rain-filled tire to Krom, and set it gently before her. She made no gesture of gratitude.

As with the complex pattern of behavior exhibited by Little Bee, it is very hard to suppose Jakie's action stemmed from the operation of some automatic precognitive mechanism. The whimsical hypothesis that, as he saw Krom's efforts, his own mirror neurons fired in ways producing a readiness for tire-pulling behavior, expressed in imitation of her efforts, could only beguile us if we ignored the direction of his actions toward the release of the *innermost* tire and the subsequent *careful* carrying of that tire *to Krom*. To explain what he did, we must credit him with recognizing that Krom wanted the innermost tire—with the water inside it.

25. Frans de Waal, *Good Natured* (Cambridge, MA: Harvard University Press, 1996), 83.

Jakie modified his wishes from what they would have been in Krom's absence, and he did so in light of his perception of her desires. He aligned his wants with hers. Are there grounds for skepticism about his altruism? If so, they must stem from concerns that the anti-Machiavelli condition is violated. Perhaps Jakie expected some future reciprocation—but that would be to impute to him a seriously misguided appraisal of Krom's future abilities to reward him (an appraisal quite at odds with his clear social intelligence; Jakie understands Krom's place in the troop). Perhaps he aimed to impress others—but Jakie was surely aware that the only other primate around was the (socially irrelevant) de Waal. Or should we think Jakie not only feels glows and pangs, but has the cognitive powers to perceive the present causes of their future occurrence? Skeptics about altruism are often moved by the thought that an egoistic story is less extravagant than a hypothesis introducing some ability to identify with others. Here, however, skeptical hypotheses about glows and pangs seem the truly extravagant options.

So we can broaden the domain of psychological altruism in the nonhuman world, at least a little. This is important for understanding the ethical project, because it allows us to attribute altruistic desires to animals *before ethical considerations are on the scene.* A central theme of my approach to altruism is that there are preethical forms of altruism and that these are realized in animals who have not yet acquired ethical practice. Yet caution is necessary. Besides the striking—and clear—cases, there are many instances of primate behavior suggestive of altruism, in which skeptical challenges are far harder to rebut. Observations of chimpanzees and bonobos frequently inspire the interpretation that particular pairs form genuine friendships, that the mutual adjustment of behavior signals an underlying modification of preferences and intentions, prompted by recognition of the other's wants. When the apparently stable alliance breaks down, when a "friend" deserts a seemingly close ally, there are two possible reactions: one can see this as revealing that the parties were calculating all along, using one another to mutual advantage (or apparent mutual advantage); or one can suppose it exposes the previously unnoticed limits of altruism along one of the dimensions (scope) distinguished earlier. Later in this chapter, my preferred explanation of the evolution of psychological altruism will be used to support

the hypothesis that, in some of these cases at least, we find genuine altruism.

Recent studies of human behavior often suggest that altruism is far more prevalent in our own species than in our closest evolutionary relatives, attributing the difference to the power of human cultural evolution. Although this conclusion may be correct, if psychological altruism is understood as in §§3–5, it cannot be established as easily as experimenters often believe. Indeed, as I shall suggest later, experimental results taken to support the "pervasive character of human altruism" are not concerned with *psychological* altruism at all, but with *behavioral* altruism; as we shall see (§11), some of the types of behavioral altruism involved are interesting in their own right.

Participants in interactions where there are possibilities for sharing are willing to divide a pool of money with fellows, even though they have the chance to take everything for themselves, and this finding persists across cultures.[26] The behavior counts as psychological altruism only if these subjects are responding to the wants of their perceived beneficiaries and the response is also not the result of an attempt to satisfy solitary preferences. One might worry about both conditions. First, these participants have little knowledge of the wants of their beneficiaries. It is thus hard to view their response as a modification of preferences through perception of another's wants or needs. Second, the skeptical hypothesis that apparently altruistic behavior is driven by desires to achieve glows or avoid pangs has considerable plausibility in these conditions. It is hard to rule out the suggestion that these people share as they do because they want to accord with (or do not want to violate) canons of approved social behavior. They are behavioral altruists whose motivations are not readily characterized as either altruistic or egoistic.

Reflection on the experiments raises the disturbing thought that there is important kinship between the performances of these behavioral altruists and those of their counterparts in earlier studies of willingness to inflict pain and punishment—to administer electric shocks to people

26. I ignore here the variety of ways in which opportunities for sharing arise, and, in particular, the important point that subjects will sometimes give some of their assigned money to "punish" participants who fail to share. For a more extensive discussion, see §11.

who are allegedly being "trained" or to function as an effective "prison guard."[27] In both types of psychological experiments, the behavior elicited, whether apparently callous (even "monstrous") or apparently altruistic, may largely express a desire to conform to social expectations.

Perhaps the precise and imaginative experiments on sharing behavior are not really concerned with *psychological* altruism. Demonstrating the conditions for psychological altruism is demanding. One should conceive altruism as covering both the nonhuman examples discussed earlier and the behavior of the experimental subjects, without raising awkward issues about motivations. For some purposes it is surely more appropriate to concentrate on behavioral altruism—if, for example, one wants to scrutinize the hypothesis that economic agents always behave as rational self-interested agents, exploring the possibilities of behavioral altruism is exactly what is needed.

For our purposes, however, there are two reasons to focus on the more demanding notion of psychological altruism. Those who recognize and respond to the wishes of others are different in important ways from people who are moved to help solely by their desire to be well regarded or to have the narcissistic comforts of self-congratulation. The conjecture that similar motivations pervade the studies of sharing behavior and of willingness to torture brings home the point in a dramatic way—even though we might not want to lump the sharers with those who administered "shocks" in the "very dangerous" range, the recognition of an underlying propensity to conform in both situations reminds us that aiming at conformity can blind one to the wants of others with damaging consequences.

More important, if one hopes to understand how ethical practices grew out of human capacities for psychological altruism, the *conception of psychological altruism will have to be prior to that of behavior done in accordance with, or out of regard for, social norms or ethical maxims.* If, as seems likely, the actions of many of the experimental subjects express their wish to exemplify norms of sharing, then their "altruism," if we call it that, will be a product of their immersion in the ethical practice of their

27. For a concise and informative survey of these experiments, see John Sabini and Maury Silver, *Moralities of Everyday Life* (Oxford, UK: Oxford University Press, 1982), chap. 4.

community. Behavioral altruism of this sort cannot be found in the societies in which the ethical project began. We cannot trace the project to prior dispositions to altruism, if we suppose that the prior dispositions are forms of behavioral altruism grounded in acceptance of ethical maxims.

My persistence in advocating a demanding conception of psychological altruism allows for interesting and valuable forms of human action besides the psychologically altruistic ones. Altruism, to repeat, is a complex notion. As we shall discover later, the taxonomy of human action has further complications—it would be wrong to suppose that everything else, besides psychological altruism, is undifferentiatedly and brutishly selfish. In understanding the evolution of human ethical practice, further distinctions and conceptions will be needed (see §11); at that stage it will be possible to provide a more adequate view of the experimental research alluded to here.

For the time being, it suffices to acknowledge some examples of psychological altruism, manifested in other primate species and in our own, besides the fundamental instances of maternal concern. The next step will be to understand how altruistic dispositions might have originated and been maintained under natural selection. We turn to the second part of the task assigned at the beginning of this chapter: to show that dispositions to psychological altruism were necessary for the type of society shared by our hominid ancestors, chimpanzees, and bonobos.

§8. Possibilities of Evolutionary Explanation

The most fundamental forms of psychological altruism, concern for offspring and, more broadly, altruistic tendencies toward close relatives, can readily be understood in terms of kin selection (as already indicated in §6). If an organism tends to adjust its preferences in response to the perceived wants of others (in accordance with the conditions of §3), if there is an allele (or alleles) that underlies that tendency, if the others who benefit from the tendency are relatives, and if the extent of the benefit is sufficiently large with respect to the personal sacrifices (gauged in terms of reproductive success) made by the altruistic animal, the allele(s) and the tendency will spread under natural selection.[28] Kin selection

28. For details, see Kitcher, "Evolution of Human Altruism."

allows for psychological altruism as *one* mechanism for helping behavior toward relatives, but it will equally favor *any* mechanism achieving the same effects. The fact that psychological altruism issuing in aid toward relatives would have been favored by kin selection does not entail that it must therefore exist. In §6, psychological altruism was defended as the best explanation for some types of sharing and helping behavior toward young. *Given* the altruistic tendency, kin selection is the most likely explanation of its presence. (Of course, it would count against the original attribution of psychological altruism to primate mothers if there were no plausible evolutionary explanation.)

Section 7 began with the poignant example of Little Bee and her mother, and here too there is a ready explanation in terms of kin selection. Imagine an original state in which the only form of psychological altruism is directed toward offspring. Suppose a new variant arises, a genetic change causes (in the pertinent environment) a tendency to broaden the range of altruism, allowing for possibilities that other animals, besides the young, will provoke that modification of preferences constitutive of psychological altruism. Animals with the variant are less fussy about those they want to help, but their altruistic responses are always toward close relatives. For concreteness, assume that an animal with the variant has the original tendency to respond, when a parent, to the perceived needs of the young, as well as other tendencies to respond to the perceived needs of parents and siblings. Helping siblings and parents (although not to the same intensity with which aid is channeled toward one's own young) contributes to the spread of the variant allele: for siblings have chances to produce offspring with that allele, and parents likewise have opportunities for generating further young of the new type. Hence the broadening of the psychological altruism originally focused in maternal concern can be favored by kin selection.

The evolutionary scenario just outlined will account for behavior like Little Bee's. A tendency to respond to the perceived wants and needs of one's mother would be favored by kin selection, for, frequently, the helping behavior produced by the altruistic tendency would increase the mother's expected reproductive success and the frequency of the allele(s) underlying the broadening of psychological altruism. Sometimes, however, animals with the tendency may make sacrifices that far outweigh any expected returns—as exemplified by Little Bee's devotion to Madam

Bee. If their helping behavior were based on calculation, it would be grotesquely misguided, belying the animals' manifest intelligence. It is better viewed as a noncalculational, emotional response, of a type that normally increases inclusive fitness, but that, in the case at hand, has negative effects on the spread of the underlying alleles. (Madam Bee's predicament arouses altruistic emotions in Little Bee—and the disposition to be aroused in this way is generally adaptive; the altruistic emotions give rise to particular altruistic desires; on this occasion, acting on those altruistic desires detracts from reproductive success.)

Will the envisaged evolutionary account extend to the example of Jakie and Krom? Perhaps. Here the relationship is far more distant, but the sacrifice made by Jakie is also quite trivial in comparison with Little Bee's months-long dedication. A tendency to (mild) psychological altruism toward any member of the ambient social group might be favored by kin selection, for there is always a (significant?) chance it will direct aid toward relatives, and thus favor the spread of the relevant allele(s). Any hypothesis along these lines would have to be carefully elaborated—for the reproductive costs and benefits are by no means as easy to assess as in the simpler examples involving close relatives—and it also presupposes that evolution of the traits underlying primate social life can be understood prior to accounting for the spread of psychologically altruistic tendencies to group members. Animals *without* the broader tendencies would have to be able to evolve capacities for group life, so that, with the group in place, the stage would be set for kin selection to favor the expansion of psychological altruism across a broader range. In §9 I shall directly question this presupposition and argue that psychological altruism is fundamental to primate social life.

Kin selection is only one of the two mechanisms whose recognition resolved the long-standing puzzle of the evolution of *biological* altruism. The other is the disposition to reciprocate. Tendencies to engage in a pattern of interaction with other organisms, in which each participant gives up something on one occasion and reaps a greater gain in some subsequent encounter, can evolve, thus accounting for cooperation among nonrelatives.[29] The initial thought is simple and elegant. If two animals

29. This approach stems from the important work of Robert Trivers, William Hamilton, and Robert Axelrod. The Trivers-Hamilton-Axelrod approach has given rise to an

share a propensity for making small sacrifices (measured in terms of reproductive success) to promote greater (reproductive) benefits for the other, if they interact repeatedly, and if the propensity has a genetic basis, then each may reap (reproductive) advantages from the sequence of interactions. Suppose you and I are the animals in question. Today I help you to some significant biological benefit, at much smaller reproductive cost to myself. Tomorrow, you return the favor. Each of us has made a net gain (measured in terms of reproductive prospects). The longer we continue, the larger the benefits we garner. The apparently pedantic introduction of qualifying terms—"biological," "reproductive"—is important because a mode of evolutionary explanation for *biological* altruism does not automatically provide a convincing account of the evolution of *psychological* altruism. With respect to kin selection, the situation is different, for kin selection is neutral in regard to whether psychological altruism underlies the pertinent forms of helping behavior: where one can argue that psychological altruism is the best explanation of that behavior (as with the case of maternal concern; see §6), viewing the tendency as the product of kin selection does nothing to undermine the argument or its conclusion. Reciprocal altruism, by contrast, precisely because of the simplicity of the idea, invites the skeptical complaint that calculational mechanisms are at work, and that the anti-Machiavelli condition is violated. To put the point bluntly, whenever a tendency to a form of behavior can evolve through reciprocal altruism, it looks as though animals with the cognitive sophistication required for psychological altruism would also have the abilities to make a calculation revealing how the behavioral propensity would satisfy their own solitary preferences; hence there would be grounds for skepticism about any alleged psychological altruism. At the very least, when tendencies to behavior are explained by supposing they evolved through reciprocal altruism, skeptics seem to have a forceful objection to the attribution of

---

extensive series of further investigations. See, for example, Alexander Harcourt and Frans de Waal, *Coalitions and Alliances in Humans and Other Animals* (Oxford, UK: Oxford University Press, 1992); Karl Sigmund, *Games of Life* (New York: Oxford University Press, 1993); and Ronald Noë, Jan van Hoff, and Peter Hammerstein, eds., *Economics in Nature* (Cambridge, UK: Cambridge University Press, 2001).

psychological altruism: the animals can identify the long-term advantages of trading favors.

If reciprocal altruism were the fundamental mechanism through which cooperative behavior between unrelated animals evolved, we should have to meet this concern directly, showing that genuine psychological altruism could emerge and be maintained because of the (reproductive) advantages of reciprocation. I shall proceed differently. Patterns of reciprocation have to rest on something more basic, tacitly assumed by accounts of reciprocal altruism. This more basic evolutionary mechanism favors the emergence of tendencies to psychological altruism. Let us start by reviewing how cooperation among unrelated animals is typically explored.

Interactions among animals can be seen as games, in which the players pursue "strategies" (of which they may or may not be conscious). The outcomes of each combination of strategies are represented by the "payoffs" to the players, assignments of numbers representing the values for them of what occurs (for evolutionary studies, these values are the effects on their reproductive success). Evolutionary game theory approaches reciprocation among nonrelatives by considering games involving possibilities of cooperation and also of competition. One particular game has received great attention, the famous prisoner's dilemma (PD).

In PD, each player has two options: to cooperate or to defect. If one cooperates and the other defects, the former obtains the *sucker's payoff,* while the latter enjoys the *traitor's payoff*. If both cooperate, they reap the *reward for mutual cooperation*. If both defect, they both receive the *punishment for mutual defection*. A table shows the outcomes for both players (with returns to the "Row Player" listed first, and returns to the "Column Player" given second).

	C(ooperate)	D(efect)
C	$<R, R>$	$<S, T>$
D	$<T, S>$	$<P, P>$

(Here T is the traitor's payoff, R the reward for mutual cooperation, P the punishment for mutual defection, and S the sucker's payoff.) It is

supposed that $T > R > P > S$, and that $T + S < 2R$.[30] If the game is played just once, defection (D) is a dominant strategy for both players, since $T > R$ and $P > S$. Rational actors in a socioeconomic interaction of this form are expected to wind up with the noncooperative outcome of mutual punishment, rather than achieving the reward for mutual cooperation—which, if they could be assured of it, they would prefer (since $R > P$). By the same token, if animals sometimes engage in interactions with nonrelatives, where the payoffs in units of reproductive success meet the conditions of PD, natural selection would apparently favor strategies of defection.

Not, however, if the interactions are repeated. In an iterated prisoner's dilemma (IPD), players can adjust their strategies to the previous performance of those with whom they interact. A strategy for IPD consists in a choice of how to play on the first round, together with a set of preferred responses to the various potential sequences of choices by one's partner/opponent. Suppose you know the interaction will be repeated but do not know exactly how many times it will occur.[31] Your strategy is specified by saying how you will begin, and how you will act given any potential history of choices by your partner.

Robert Axelrod investigated the success of various strategies empirically, by inviting scholars to submit their preferred proposals for playing IPD, and staging a computer tournament. In each round of the tournament, different strategies were paired (as in a round-robin), and then played a particular version of PD against each other for a large number of iterations.[32] The winner was one of the simplest strategies submitted, tit for tat (TFT), which begins by cooperating, an-

30. The second condition implies that, if the game is repeated, it is cooperatively better for the players both to play *C* than to adopt a pattern of alternating *C* and *D* (so that, on each occasion, one plays sucker and the other plays traitor, with alternation of roles).

31. This last stipulation is added to address the concern that it will always be preferable to defect on the last round, that once that is a matter of common knowledge it will be rational to defect on the penultimate round, and so on. There are complications here that I shall not explore. For present purposes, it is enough to follow the standard treatment.

32. For details, see Axelrod, *Evolution of Cooperation*. Note that the number of iterations is close to two hundred, and that the payoffs in the game—the values of *T, R, P,* and *S*—are the same in each iteration and in each round.

swers defection with defection, and responds to cooperation with cooperation. In the common parlance, TFT is "nice, provokable, and forgiving."

Mathematical analyses of populations consisting of variant strategies for playing IPD suggested that TFT is *evolutionarily stable;* that is, in a population in which it is prevalent, it resists invasion by alternatives arising at low frequencies.[33] The analyses accounted for the *maintenance* of cooperative behavior under natural selection, once it has become common, but did not explain how such behavior might *originate,* evolving from an initial state in which it was rare. Unless they are strongly disposed to interact with one another rather than with the rest of the population, TFT variants, arising at low frequencies in groups full of noncooperators, are driven out by natural selection.

Two problems have now emerged with the hypothesis that psychological altruism toward nonrelatives (or psychological altruism more intense than would have been favored by kin selection acting alone) might have evolved through reciprocal altruism. First, while reciprocal altruism may help us understand cooperation, its amenability to predictive calculation raises skeptical doubts about psychological altruism as a mechanism for the cooperation. Second, it is hard to understand how dispositions to cooperate (Machiavellian or altruistic) could have obtained a first firm foothold. There is a third difficulty, too. The IPD scenario imposes very particular conditions: two animals are designated as partners for a long sequence of PDs with exactly the same structure; at the end of this, they are released and assigned to different partners for a repetition of that sequence of interactions. The idea that anything like this happened

33. For the important notion of evolutionary stability (of an *evolutionarily stable strategy*), see John Maynard Smith, *Evolution and the Theory of Games* (Cambridge, UK: Cambridge University Press, 1982). From the beginning it was apparent that there were indirect ways in which populations of TFTs could be invaded. In such populations, variants that invariably cooperate would be indistinguishable from the TFTs and could thus enter. Once there were sufficiently many of them, the stage would be set for noncooperative strategies to invade through exploiting the undifferentiating cooperators. (See my discussion in *Vaulting Ambition*, [Cambridge, MA: The MIT Press, 1985], 100–101.) Further research revealed that combinations of noncooperative strategies can also invade (Robert Boyd and J. P. Lorberbaum, "No Strategy Is Stable in the Repeated Prisoner's Dilemma," *Nature* 327 [1987]: 58–59).

in our primate past is immensely implausible. Surely no giant hand swooped down on the savannah, locking animals into compelled interactions that recapitulated the same form.

To address the difficulties with reciprocal altruism, start with the last. Far more realistic is a different scenario. Suppose our primate ancestors had recurrent opportunities for interacting with a conspecific, and, on these occasions, they could either engage in that interaction or act by themselves. Assume, too, they could sometimes choose partners for interaction, signaling their willingness (or reluctance) to engage in joint activity. This would replace the standard structure of the IPD, the repeated *compulsory* games, with something different—repeated opportunities for *optional* games (as I shall call them). The framework of optional games is both more realistic and resolves some difficulties besetting the orthodox understanding of reciprocal altruism.

An example helps to fix ideas. Our primate ancestors had to remove parasites from their fur. The task was undertaken repeatedly and could be done in either of two ways. One possibility is self-cleaning—although that poses problems because it is hard to reach some parts of the body. Another is to team up with a partner—but that risks exploitation; after the first animal has provided a thorough cleaning, the second may provide something superficial and then go off to more interesting activities. Primates could have signaled to one another their willingness to engage, issuing, accepting, and turning down invitations, so that partners for interaction could be chosen.

With some plausible assumptions about the benefits of hygiene and the costs of spending time, it can be shown that the scenario envisaged has the structure of an optional PD. If two animals interact with each other, the cooperative strategy is for each to provide a thorough grooming for the other; defecting consists in being quick and sloppy. The best of all outcomes is to receive the thorough attention of one's partner and to provide little in return; slightly less good is to obtain a serious cleaning and to return the favor; significantly less good is to receive a superficial grooming and to give back the same; even worse (although not much worse) is to clean one's partner conscientiously but receive a superficial grooming. Not interacting, "opting out," and cleaning oneself, is intermediate between mutual cooperation and mutual defection. Hence, with a some-

what arbitrary assignment of numbers, the structure of the interaction is as follows:

	C	D
C	<9,9>	<0,10>
D	<10,0>	<1,1>

Interact →

Opt out → 5

Mathematical analyses reveal that high levels of cooperation are likely to develop, and to be sustained, in populations whose members have a sufficiently large number of opportunities for playing optional PD with one another. More exactly, a strategy of *discriminating cooperation* (DC) can originate and be maintained under natural selection.[34] Discriminating cooperators are prepared to interact with any animal that has not previously defected on them; if their only opportunities for interaction involve partners who have previously defected on them, they opt out; whenever they interact, they cooperate. Suppose we begin with a population of *antisocial* animals, beings who interact and defect with one another. In this state *asociality* will be favored: the solo strategy (always opt out) does better. In an asocial population (full of solos), however, a lone DC does equally well; there are no opportunities for interacting, and DCs are left partnerless to behave like solos. Once a second DC is present, however, the two of them team up for a happy life of cooperative interactions that bring large advantages over their asocial fellows. So, from antisociality, the population proceeds via asociality to a state of high levels of cooperation. Those high levels will be maintained until the frequency of *non*discriminating cooperators becomes sufficiently high (among DCs, nondiscriminating cooperators are invisible—they are never exploited) to allow antisocial types to enter and take advantage of them. When that happens, the population can relapse to an antisocial state.

34. The results summarized here were originally presented in Kitcher, "Evolution of Human Altruism." I should note that the strategy DC described here is characterized as DA in the earlier paper ("discriminating cooperator" is a more accurate label than "discriminating altruist").

Computer simulations reveal that the history of high levels of cooperation is quite long.[35]

There are further encouraging results about the mechanisms of cooperation. Suppose we abstract from some of the conditions I placed on psychological altruism in §3, and, in particular, from the Machiavellian concerns about calculation. Let *quasi altruists* be individuals who meet conditions 1–3 but not necessarily condition 4: they adjust their preferences to align them more closely with what they take to be the wishes of others, but they may do so on the basis of considerations of their own expected narrow benefit. As in the discussions of §§3 and 5, it is possible to gauge the intensity of the quasi altruist's response, in terms of the weight assigned to the perceived wishes of the other. Under a regime of repeated opportunities for playing optional games of various types, selection will favor quasi altruism of a more intense kind, up to golden-rule quasi altruism because quasi altruists with more intense responses will participate in a broader class of profitable interactions with others.[36]

Replacing the scenario of compulsory IPD with the framework of optional games helps. Not only does it offer a more realistic scenario for the evolution of cooperation, but it overcomes the problem of understanding how cooperation got going. It even points toward some conclusions about the mechanisms underlying cooperation: selection will favor tendencies to respond to the wants of others that give the others' preferences as much weight as one's own. Plainly, however, the shift does not address the most fundamental difficulty in using reciprocal altruism to explain the evolution of psychological altruism—*for it preserves the simplicity that invites skepticism.* Animals with the cognitive resources to count as psychological altruists would be able to see the advantages of discriminating cooperation and of being prepared to cooperate across a wide range of types of interaction. The scenario thus shows how Machiavellian calculators might have evolved to behave like golden-rule altruists.

To address this problem, to show how full-fledged psychological altruism of kinds going beyond those favored by kin selection might have

35. See John Batali and Philip Kitcher, "Evolution of Altruism in Optional and Compulsory Games," *Journal of Theoretical Biology* 175 (1995): 161–71.

36. This result is derived in Kitcher, "Evolution of Human Altruism." Note that *quasi* altruists resemble *behavioral* altruists, although some behavioral altruists may not meet condition 3 of §3.

evolved, requires a more decisive break with the mechanism of reciprocal altruism. Analyses in terms of both compulsory and optional games can play a role in understanding human social practice. The evolution of primate sociality, however, is based on a different scenario, one favoring the emergence of psychological altruism.

For optional games presuppose certain forms of cooperative abilities that have not yet been explained.

§9. The Coalition Game

Worries about the realism of the scenarios so far envisaged should remain. The primatological work of the past decades queries some assumptions hidden behind the mathematical analyses. Assuming our evolutionary cousins serve as good models of our primate pasts, can we really suppose our ancestors behaved like discriminating cooperators? On the one hand, chimpanzees and bonobos seem not to cooperate anywhere near as much as the conception of them as discriminating cooperators suggests. Moreover, they often fail to cooperate with the "right" partners—in joint hunting, for example, those who help bring down the prey are not always rewarded, while those who have not taken part end up with pieces of the spoils, and yet the dispossessed appear willing to return the next day for a similar expedition.[37] More generally, chimpanzee and bonobo societies are pervaded by asymmetries the account fails to recognize. Grooming partnerships embody some of these asymmetries, and a more focused look at grooming shows it to be a far more complicated phenomenon than the analysis outlined in §8 pretended. If considerations of hygiene alone were pertinent, it would be impossible to understand the enormous amounts of time chimpanzees and bonobos devote to grooming one another. During some periods in the recorded histories of primate troops, particularly when social tensions are running high, the animals devote three to six hours per day to plucking and smoothing one another's fur.[38]

37. See, for example, Goodall, *Chimpanzees of Gombe*, 288–89.

38. See Frans de Waal, *Chimpanzee Politics* (Baltimore: Johns Hopkins University Press, 1984), and *Peacemaking Among Primates* (Cambridge, MA: Harvard University Press, 1989).

These features of primate societies point to the more fundamental presupposition of the explanations in terms of reciprocal altruism: *these are animals who can endure one another's presence, who can occupy the same region together at the same time.* In the original IPD scenario, that is simply achieved by *force majeure;* the organisms are locked together in their long sequence of interactions. Although the shift to optional games increased the realism, it took for granted the existence of a pool of potential partners. Animals were supposed to encounter others quite frequently and to be able to signal their willingness to interact. For that, a minimal form of sociality must already be in place—the animals must be sufficiently tolerant of one another's presence to form the pool. Reciprocal altruism presupposes an ability to treat others as potential partners and not as dangerous rivals.

That ability should be the first and fundamental target of evolutionary explanation. The processes that gave rise to it generated a capacity for psychological altruism of a more extensive type than those understood in §8 in terms of kin selection.

Begin with some well-established conclusions about social life among the apes. Within this relatively small group, the extent to which social relations, tolerance, and cooperation extend beyond the family varies greatly. Gibbons divide into small family groups (mother, father, and young) that are typically hostile to outsiders. Male orangutans are mostly solitary, ready to defend their territories against incursions from other males; they interact only perfunctorily with the females whose home ranges lie within those territories; the extent of female-female association is a matter of controversy (with older orthodoxy supposing that females travel with one or two offspring, and newer observations pointing to intermittent joining of pairs of females). Groups of gorillas typically contain several adult females but have only one adult male; to a first approximation, gorilla social life involves some cooperation among unrelated females and only aggressive interactions among adult males.[39] For larger social units, with cooperation among unrelated adults of both

39. A valuable source for discussions of social life among the apes is Barbara Smuts; Dorothy Cheney, Robert Seyfarth, Richard Wrangham, and Thomas Struhsaker eds., *Primate Societies* (Chicago: University of Chicago Press, 1987).

sexes, we must turn to our evolutionary cousins, chimpanzees and bonobos.

Chimpanzees live in bisexual groups (varying in size from about 20 to approximately 100), within which there are shifting patterns of alliances and dominance relations. Among bonobos, the groups are somewhat larger (roughly 50 to 150), with the same sorts of changing internal structures.[40] A principal difference between the two groups is that the major associations in the wild seem to be among chimpanzee males and among bonobo females, although in both species, there are important social interactions among members of the other sex (and between members of opposite sexes). Study of hominid remains suggests that our ancestors lived in mixed groups and that their size was of the same order as that found in living chimpanzees and bonobos. How did the chimp-bonobo-hominid pattern of sociality evolve?

Any answer to the question must identify the features that distinguish chimps and bonobos from the other great apes. I shall develop an approach originally outlined by Richard Wrangham, who proposed that female behavior is shaped directly by ecological factors, particularly the distribution of the foods consumed by the species; males have to adapt to this distribution, adjusting their behavior to increase the chances of copulating with estrous females.[41] Crucial for our purposes is the conjecture

40. I shall tend to take chimpanzees, rather than bonobos, as the model for our hominid past. This decision rests partly on a sense that many small human societies that live in environmental conditions closer to those of our ancestors appear to share the relative intolerance for neighbors that is so marked in chimpanzee social life, and, more important, on the hypothesis that psychologically altruistic tendencies are more prominent and pervasive in bonobos than in the (common) chimpanzee. Hence I assume that if a compelling story about the evolution of sociality and its roots in psychological altruism can be given for chimpanzees, it would be easier to defend a similar account for bonobos. (Here I am indebted to a valuable conversation with Frans de Waal.)

41. See Richard Wrangham, "On the Evolution of Ape Social Systems," *Social Science Information* 18 (1979): 334–68; "An Ecological Model of Female-Bonded Primate Groups," *Behaviour* 75 (1980): 262–300); "Social Relationships in Comparative Perspective," in *Primate Social Relationships: An Integrated Approach,* ed. Robert Hinde (Oxford: Blackwell, 1983); and "Evolution of Social Structure," in Smuts et al., *Primate Societies*, 282–96. Wrangham bases his analysis on the hypothesis that the principal determinant of female reproductive success will be her access to food and that the principal determinant of male reproductive success will be the ability to copulate as frequently as possible with

that mutually hostile communities of chimpanzees have "evolved from a hypothetical solitary-male system because males could afford to travel in small parties, even though the optimal foraging strategy was to travel alone; they were forced to do so because lone males therefore became vulnerable to attacks by pairs."[42] Abstracting from the emphasis on foraging, one may recognize that, in a world with scarce resources—of whatever kind—competition among vulnerable animals may require their participation in coalitions and alliances. Addressing that problem is prior to realizing possibilities for cooperation: *for understanding cooperative interactions among unrelated animals, PD (whether optional or compulsory) is not fundamental; the framework for the games animals play is set by the problem of forming coalitions and alliances.*[43]

Imagine a population of solitary organisms (the largest units being mothers with dependent young) in an environment in which each must obtain a certain number of resources in order to survive and reproduce. Suppose the resources are scarce, the animals fight over these resources, and the stronger typically win. A five-stage process could have led from the initial situation—no cooperation except for maternal care in early life—to the kind of social structure found in chimpanzees, bonobos, and

---

estrous females. So, for example, on his account, orangutans pursue their relatively solitary lives because females can most efficiently forage for fruit by working alone, and males have physical abilities to defend a territory including the smaller home ranges of several females. I shall make no such specific assumptions. Instead, I abstract from the particularities of Wrangham's discussion, offering a more general model of which his approach would be a special case.

42. Wrangham, "Evolution of Social Structure," 290. Compare Hobbes: ". . . the weakest has strength enough to kill the strongest, either by secret machination, or by confederation with others that are in the same danger with himself" (*Leviathan,* 82). Hobbes, however, would not have thought that this could apply to the brutes, because, without speech ". . . there had been amongst men neither Common-wealth, nor Society, nor Contract, nor Peace, no more than amongst Lyons, Bears, and Wolves" (*Leviathan,* 20). Hobbes underrated the lions and the wolves and knew nothing of the chimpanzees and the bonobos [New York: Oxford University Press (World's Classics) 2008].

43. Some primatologists have recognized the point in the context of their studies of particular societies. See, for example, R. Noë, "Alliance Formation Among Male Baboons: Shopping for Profitable Partners," in *Coalitions and Alliances in Humans and Other Animals,* ed. A. Harcourt and F. de Waal (Oxford, UK: Oxford University Press, 1992), 285–321.

hominids. (Note that what is required here is an account of how a form of social structure we independently know to exist could have emerged and remained stable under natural selection: a "how possibly" explanation.)

1. **Asociality**—animals range alone (at most accompanied by dependent young), finding some resources without contest ("scramble" competition) and competing directly for others ("contest" competition).
2. **First Coalitions**—some animals arise that are disposed to act together in contest and to share the resources obtained (not necessarily equally).
3. **Escalation**—because of the success of the early coalitions, larger coalitions form, sharing the benefits they earn in contests (not necessarily equally, and possibly involving interactions among subcoalitions).
4. **Community Stabilization**—coalition size is ultimately limited by the difficulty of defending all the resources in a range, and the habitat becomes partitioned into ranges defended by stable communities, within which the resources are divided by the formation of subcoalitions.
5. **Cooperation**—by engaging in optional games (some of which may be optional PD) and behaving cooperatively, members of the stable communities increase their fitness.[44]

Without pursuing the technical details, I shall try to show how this process might unfold.

Begin with a more benign version of the initial state, a Rousseauian world that contains more than enough for everyone. As the population expands, competition enters. Eventually, so long as the competition goes on in the assumed way, some animals will not find the resources they need to survive.

44. Note that the fitness values that occur in the payoff matrices for the games played by community members, whether optional or compulsory, must reflect the consequences of actions for the underlying alliances to which the animals belong. This recapitulates the point made earlier that the structure of animal interaction cannot be understood in isolation from the demands of the most fundamental game, here seen as the coalition game.

If the animals pursue solitary strategies for gaining resources, as envisaged in stage 1 of the process outlined above, there will be contests for some resources. Assume, for simplicity, that the contests are resolved without actual fighting: the animals simply assess one another's strength, and the weaker one retires (in cases of equal strength, divisible resources are shared equally; indivisible resources are assigned to each animal with probability $^1/_2$). Throughout the course of their lives, the strength of the animals changes, according to an obvious schedule. Initially, while an animal is under the protection of its mother, it effectively has whatever strength its mother has. Once released from its mother's care, it is at its weakest. Thereafter, strength grows as the animal matures, provided that sufficient resources are obtained; eventually, perhaps, animals that live long enough undergo a slight decline in strength.

Populations faced with these conditions are vulnerable to extinction. For a new generation to arise, the young must survive the critical period after their release from maternal care. During this period, they are the weakest members of the population, and whatever they achieve must be gained by finding resources currently uncontested by others and consuming those resources before a stronger individual arrives to dispossess them. If the competition is sufficiently severe, all resources will be contested, and, after a brief period of maternal care, all the young members of the population die. In a very hostile world, populations stuck at stage 1 are likely to be short-lived. More exactly, the pressure of mortality will cull the population so it is effectively returned to a more benign—Rousseauian—environment.

Suppose, however, variants arise that are disposed to team up with others. Specifically, imagine a variant that is prepared, when weak, to travel around with another animal of similar weakness, to collect resources together and to divide them. (There may be variation in the propensities to tolerate different schedules of division.) If two such variants encounter each other, they form a coalition. Because the members of the coalition have to travel together, the coalition can visit only as many resources as a single individual can. Assume that strength is additive; that is, the strength of a coalition is the sum of the strengths of its members. Each of the variants in a coalitional pair can now increase its access to resources, for the doubled strength will surely provide victory in contests with other weak young animals and may be sufficient to win some encounters with older members of

the population. Selection thus favors variants of this type, even if the divisions of the resources acquired are not even.

Plainly, several parameters must be set in developing versions of the scenario I am envisaging, but it is possible to show that, given almost all ways of choosing values for non-Rousseauian worlds, any population at stage 1 will contain at least one pair of organisms who can increase their fitness through coalition formation. That does not mean, of course, that the disposition to team up *must* evolve: there might be no way to generate any such propensity. I shall suggest shortly that more basic capacities for psychological altruism provide a way in which the successful variants might emerge.

Just as stage 1 would favor the emergence of pairwise coalitions, so too the emergence of pairs puts pressure on animals who are working alone. The gains of the animals who team up are obtained by dispossessing those who would otherwise have done better. Any variation that equips them with a disposition to pair with another animal will be favored. As the population becomes full of coalitional pairs competing with one another, the weakest pairs will do better if they are prepared to add single members or merge with other pairs. Selection favors the variants who unite with others at the size required by the actual escalation of coalition formation.

Although the origination and escalation of coalition formation is easy to understand, the termination of the process appears more mysterious. The rationale, however, is a direct consequence of the fact that coalitions have to travel together if they are to exert their joint power. No coalition can visit more resources than a single individual. When the environment is filled with large coalitions, coalition members who receive the smallest shares may have no better option than to resume scrambling for resources the large coalitions are not able to visit. The dynamics of the process leads to a situation in which the habitat is partitioned into territories controlled by sizable coalitions, occasionally with a floating population of individuals who live on the fringes.[45]

45. The announced results are not hard to derive analytically. They coincide with the findings of some ingenious computer simulations designed by Dr. Herbert Roseman. See his unpublished Ph.D. dissertation "Altruism, Evolution, and Optional Games,"2008 Columbia University.

This is an evolutionary scenario for the emergence of the social structure found in chimpanzees, bonobos, and hominids, one that will lead to groups mixed by age and sex, each of which controls a relatively stable territory that it defends against neighboring rivals. Within these groups, there will be patterns of alliances bearing on the division of the resources the group commands. That structure will determine the potential benefits of various possibilities for cooperation, for there will be gains from strengthening existing alliances and costs from disrupting them. The homogeneous pool of partners for optional games, envisaged in the analysis of §8, is structured by the shapes of previous encounters. Reciprocal altruism and interaction in optional games can be understood only against the background of the coalitional structure of the group.

So far the conclusions address only animal behavior, with no direct implications about psychological altruism. To go further, it is necessary to ask how the variants envisaged, with their disposition to team up with others, might have been psychologically realized. Answer: this ability to form coalitions, and ultimately to constitute a stable social group, expresses a further expansion of those fundamental psychologically altruistic tendencies attributed in the case of maternal care.

Mothers have a propensity to modify their wants and preferences from what they would otherwise have been, to accommodate the perceived wants of the young. Primates have evolved to broaden this response to others, so that preferences reflect the perceived wishes of close relatives, a broadening supported by kin selection and manifest in the behavior of Little Bee. I propose a further extension: the disposition to adjust wants and preferences to the perceived preferences of an age-mate, initially triggered in contexts where both animals are weak and vulnerable. This is a species of psychological altruism, the capacity for early friendship. Pairs of animals with this broadened altruistic disposition reap the advantages just outlined. Young animals, no longer under parental protection, need allies if they are to gain anything in a competitive world. Psychological altruism of this special type is one way for them to find friends.

Skeptics will suppose there are self-interested routes to the same end. What would they be? The coalition game is by no means a simple opportunity for reciprocal altruism. It does not present the players with a compelled or optional iterated prisoner's dilemma, inviting them to cal-

culate a strategy for success. The coalition game is many-personal—and, for the players, the number of participants will typically be unknown. It is not even evident what would count as a "best strategy" for playing it. Whether someone counts as a good ally or not depends on all sorts of delicate facts animals have no way of recognizing. Moreover, working out a good procedure for playing the game challenges the intellectual powers of mathematicians, economists, and philosophers. The best one can do is pick a partner, team up—and hope.

That appears to be just what chimpanzees and bonobos do. Their alliances do not seem to depend on any tallying of costs and benefits. Instead, these animals are prepared to support members of their groups with whom they have a history of interactions, often dating back to periods early in their lives—the strongest alliances descend from that period of juvenile vulnerability.[46] What sorts of calculations might underlie their behavior?

It is natural to believe that the clever head can always substitute for the kindly heart, but that need not always be so. When the problems posed for reasoned selection of the best strategy are sufficiently intractable—as they are in the case of the coalition game—it may not just be that an emotional response to another animal, the transfer of altruistic dispositions to identify with others to a novel sphere, the domain of "early friendship," does no worse than the cunning of the Machiavellian calculator, but that it works *better.* Animals with a disposition to try to work out the costs and benefits suffer from too little information to make good decisions on this basis, and their efforts can easily lead them to abandon an alliance when there are no serious prospects for doing any better. Furthermore, they may hesitate more than their blindly sympathetic counterparts, and indeed be recognizable by others as less reliable and less stable coalition partners.

When weak animals are forced to compete for resources they need, their inability to win contests by themselves confers a selective advantage on a disposition to identify with the interests of conspecifics, particularly with those who are in a similar predicament. That advantage fostered the spread of propensities to psychological altruism antecedently limited,

46. See Goodall, *Chimpanzees of Gombe*, 379–85, 418–24.

first toward young and then toward close relatives. The broadened propensities allowed for the formation of those loose coalitions found in our evolutionary cousins. Far from being anthropomorphic, sentimental, or self-deceiving, the hypothesis advanced here looks like the best explanation of the form of sociality of our hominid past. It also explains why the friendships of youth are so deep and enduring, both in human beings and in other primates, and why newcomers are sometimes accepted into primate social groups when a resident animal has formed social bonds with them in a shared past as juveniles together.[47]

Psychological altruism is the kernel from which ethical practice grows—because it lies at the heart of the type of sociality our hominid ancestors experienced. As we shall discover, however, the plant is far more elaborate than the seed.

47. De Waal relates a striking instance, in which a relatively unprepossessing male (Jimoh) was accepted into a chimpanzee troop because of his prior association with two older females in the group. See de Waal, *Good Natured*, 131–32.