# *Is human morality innate?*

**Richard Joyce**

Penultimate draft of the chapter appearing in P. Carruthers, S. Laurence & S. Stich (eds.),
*The Innate Mind: Culture and Cognition* (Oxford University Press: 2006): 257-279.
[A shorter version is reprinted in M. Ruse (ed.), *Philosophy after Darwin: Classic and Contemporary Readings* (Princeton University Press, 2009) 452-463.]

The first objective of this chapter is to clarify what might be meant by the claim that human morality is innate. The second is to argue that if human morality is indeed innate an explanation may be provided that does not resort to an appeal to group selection, but invokes only individual selection and so-called "reciprocal altruism" in particular. This second task is not motivated by any theoretical or methodological prejudice against group selection; I willingly concede that group selection is a legitimate evolutionary process, and that it may well have had the dominant hand in the evolution of human morality. There is a fact of the matter about which process, or which combination of processes, produced any given adaptation, and it is to be hoped that in time enough evidence might be brought into the light to settle such issues. At present, though, the evidence is insufficient regarding human morality. By preferring to focus on reciprocity rather than group selection I take myself simply to be outlining and advocating a coherent and uncomplicated hypothesis, which may then take its place alongside other hypotheses to face the tribunal of our best evidence.

## 1 Understanding the hypothesis

Before we can assess the truth of a hypothesis, we need to understand its content. What might it mean to assert that human morality is innate? First, there are issues concerning what is meant by "innate." Some have argued that the notion is so confused that it should be eliminated from serious debate (see Bateson, 1991; Griffiths, 2002). I think such pessimism is unwarranted, but I agree that anyone who uses "innate" in critical discussion should state what he or she has in mind. I suggest that what people generally mean when they debate the "innateness of morality" is whether morality (under some specification) can be given an adaptive explanation in genetic terms: whether the present-day existence of the trait is to be explained by reference to a genotype having granted ancestors reproductive advantage, rather than by reference to psychological processes of acquisition.[1] If morality is innate in this manner, it would not follow that there is a "gene for morality." Nor do this conception of innateness and the references to "human nature" that routinely come along with it imply any dubious metaphysics regarding a human essence. Asserting that bipedalism is innate and part of human nature doesn't imply that it is a necessary condition for being human.

---

Much of this chapter is a condensed version of arguments presented in *The Evolution of Morality* (MIT Press, 2006). Many passages are taken straight from this book.

[1] This stipulation is not intended as an analysis or a general explication of the concept *innateness*. I have no objection to the term's being used in a different manner in other discourses.

Nor does it follow that an innate trait will develop irrespective of the environment (for that isn't true of any phenotypic trait) or even that it is highly canalized. The question of how easily environmental factors may affect or even prevent the development of any genetically encoded trait is an empirical one that must be addressed on a case-by-case basis. It is also conceivable that the tendency to make moral judgments is the output of an innate conditional strategy, in which case even the existence of societies with nothing recognizable as a moral system would not be inconsistent with morality's being part of human nature, for such societies may not satisfy the antecedent of the conditional. Indeed, if our living conditions are sufficiently dissimilar from those of our ancestors, then, in principle, there might have been *no* modern society with a moral system—not a single moral human in the whole wide modern world—and yet the claim that morality is innate might remain defensible. These possibilities are highlighted just to emphasize the point that something's being part of our nature by no means makes its manifestation inevitable. But, of course, we know that in fact modern human societies do have moral systems; indeed, apparently all of them do (see Roberts, 1979; Brown, 1991; Rozin et al., 1999).

The hypothesis that morality is innate is not undermined by observation of the great variation in moral codes across human communities, for the claim need not be interpreted as holding that morality with some particular content is fixed in human nature. The analogous claim that humans have innate language-learning mechanisms does not imply that Japanese, Italian, or Swahili is innate. We are prepared to learn some language or other, and the social environment determines which one. Although there is no doubt that the content and the contours of any morality are highly influenced by culture, it may be that the fact that a community has a morality *at all* is to be explained by reference to dedicated psychological mechanisms forged by biological natural selection. Even if mechanisms of cultural transmission play an exhaustive role in determining the content of an individual's moral convictions, this would be consistent with there being an innate "moral sense" designed precisely to make this particular kind of cultural transmission possible. That said, it is perfectly possible that natural selection has taken *some* interest in the content of morality, perhaps favoring broad and general universals. (Later, I will mention some evidence indicating that there are a number of recurrent themes among all moral systems.) This "fixed" content would pertain to actions and judgments that enhance fitness despite the variability of ancestral environments. Flexibility is good if the environment varies; but if in some respect the environment is very stable—for example, it is hard to imagine an ongoing situation where fitness will be enhanced by eating one's children—then moral attitudes with fixed content may be more efficient. After all, speaking generally, phenotypic plasticity can be costly: Learning introduces the dangers of trial-and-error experimentation, and it takes a potentially costly amount of time. (Consider the nastiness of getting a sun burn before your skin tans in response to an increase in sun exposure, or the dangers of suffering a disease before your immune system kicks in to combat it.)

Apart from controversy surrounding innateness (which I don't for a second judge the foregoing clarifications to have settled), the hypothesis that human morality is innate is also bedeviled by obscurity concerning what might be meant by "morality." A step towards clarity is achieved if we make an important disambiguation. On the one hand, the claim that humans are naturally moral animals might mean that we naturally act in ways that are morally laudable—that the process of evolution has designed us to be social, friendly, benevolent, fair, and so on. No one who has paused to glance around herself will ever claim that humans *always* manifest such

virtuous behaviors, for it is obvious that we can also be violent, selfish, lying, insensitive, and unspeakably nasty creatures. By saying that humans naturally act in morally laudable ways, we might mean that these morally unpleasant aspects of human behavior are "unnatural," or that both aspects are innate but that the morally praiseworthy elements are predominant, or simply that there exist some morally laudable aspects among what has been given by nature, irrespective of what darker elements may also be present.

Alternatively, the hypothesis that humans are by nature moral animals may be understood in a different way: as meaning that the process of evolution has designed us to think in moral terms, that biological natural selection has conferred upon us the tendency to employ moral concepts. According to the former reading, the term "moral animal" means *an animal that is morally praiseworthy*; according to the second, it means *an animal that morally judges*. Like the former interpretation, the latter admits of variation: Saying that we naturally make moral judgments may mean that we are designed to have particular moral attitudes towards particular kinds of things (for example, finding incest and patricide morally offensive), or it may mean that we have a proclivity to find something-or-other morally offensive (morally praiseworthy, etc.), where the content is determined by contingent environmental and cultural factors. These possibilities represent ends of a continuum; thus many intermediate positions are tenable.

These two hypotheses might be logically related: It has often been argued that only beings who are motivated by moral thoughts properly deserve moral appraisal. If this relation is correct, then humans cannot be naturally morally laudable unless we are also naturally able to employ moral judgments; thus establishing the truth of the first hypothesis would suffice to establish the truth of the second. However, this strategy is not a promising one, because the connection mentioned—roughly, that moral appraisal of an individual implies that the individual is morally motivated—is too contentious to rest arguments upon with any confidence. (In fact, as I will mention below, I doubt that it is true.)

It is the second hypothesis with which this chapter is concerned, and I will be investigating it directly, not by establishing the first hypothesis. With it thus made explicit that our target hypothesis concerns whether the human capacity to make moral judgments innate, it ought to be clear that arguments and data concerning the innateness of human prosociality do not necessarily entail any conclusions about an innate morality. Bees are marvelously prosocial, but they hardly make moral judgments. An evolutionary explanation of prosocial emotions such as altruism, love, and sympathy also falls well short of being an evolutionary explanation of moral judgments. We can easily imagine a community of people, all of whom have the same desires: They all want to live in peace and harmony, and violence is unheard of. They are friendly, loving people as far as you can see, oozing with prosocial emotions. However, there is no reason to think that there is a moral judgment in sight. These imaginary beings have *inhibitions* against killing, stealing, etc.—they wouldn't dream of doing such things because they just really don't want to. But we need not credit them with a conception of a *prohibition*: the idea that one shouldn't kill or steal because it's wrong. And moral judgments require, among other things, the capacity to understand prohibitions. To refrain from doing something because you don't *want* to do it is very different from refraining from doing it because you judge that you *ought* not do it.

This point must not be confused with one famously endorsed by Immanuel Kant: that actions motivated by prosocial emotions cannot be considered morally admirable (Kant [1783] 2002, p. 199-200). I am more than happy to side with common sense against Kant on this point.

We often morally praise people whose actions are motivated by love, sympathy, and altruism. In fact, I am willing to endorse the view that on occasions a person whose motivations derive from explicit moral calculation rather than direct sympathy is manifesting a kind of moral vice. So it is not being denied that the imaginary beings described above deserve our moral praise, or even that they are, in some sense of the word, morally virtuous. My point is the far less controversial one that someone who acts solely from the motive of love or altruism *does not thereby make a moral judgment* (assuming, as seems safe, that these emotions do not necessarily involve such judgments[2]).

Now we face the question of what a moral judgment is, for we cannot profitably discuss the evolution of X unless we have a firm grasp of what X is. Unfortunately, there is disagreement among meta-ethicists, even at the most fundamental level, over this question. On this occasion I must confine myself to presenting dogmatically some plausible distinctive features of a moral judgment, without pretending to argue the case.

- Moral judgments (as public utterances) are often ways of expressing conative attitudes, such as approval, contempt, or, more generally, subscription to standards; moral judgments nevertheless also express beliefs (i.e., they are assertions).
- Moral judgments pertaining to action purport to be deliberative considerations that hold irrespective of the interests/ends of those to whom they are directed; thus they are not pieces of prudential advice.
- Moral judgments purport to be inescapable; there is no "opting out."
- Moral judgments purport to transcend human conventions.
- Moral judgments centrally govern interpersonal relations; they seem designed to combat rampant individualism in particular.
- Moral judgments imply notions of desert and justice (a system of "punishments and rewards").
- For creatures like us, the emotion of guilt (or "a moral conscience") is an important mechanism for regulating one's moral conduct.

Something to note about this list is that it includes two ways of thinking about morality: one in terms of a distinctive subject matter (concerning interpersonal relations), the other in terms of what might be called the "normative form" of morality (a particularly authoritative kind of evaluation). Both features deserve their place. A set of values governing interpersonal relations (e.g., "Killing innocents is bad") but without practical authority, which would be retracted for any person who claimed to be uninterested, for which the idea of punishing or criticizing a transgressor never arose, simply wouldn't be recognizable as a set of *moral* values. Nor would a set of binding categorical imperatives that (without any further explanation) urged one, say, to kill anybody who was mildly annoying, or to do whatever one felt like doing. (Philippa Foot once claimed that to regard a person as bad merely on the grounds that he runs round trees in a certain direction, or watches hedgehogs by the light of the moon, is not to have evaluated him

[2] Notice that my examples of prosocial emotions do not include guilt or shame, for the very reason that I accept that these emotions do involve a normative (and often moral) judgment. Guilt, I submit, necessarily involves thoughts of having transgressed.

from a *moral* point of view—it's just the wrong *kind* of thing [Foot, 1958, p. 512].) Any hypothesis concerning the evolution of a moral faculty is incomplete unless it can explain how natural selection would favor a kind of judgment with both these features.

I am not claiming that this list succeeds in capturing the necessary and sufficient conditions for moral judgments; it is doubtful that our concept of *a moral judgment* is sufficiently determinate to allow of such an exposition. Some of these items can be thought of merely as observations of features of human morality, whereas others very probably deserve the status of conceptual truths about the very nature of a moral judgment. The sensibly cautious claim to make is that so long as a kind of value system satisfies *enough* of the foregoing criteria, then it counts as a moral system. A somewhat bolder claim would be that some of the items on the list (at least one but not all) are necessary features, and enough of the remainder must be satisfied in order to have a moral judgment. In either case, how much is "enough"? It would be pointless to stipulate. The fact of the matter is determined by how we, as a linguistic population, would actually respond if faced with such a decision concerning an unfamiliar community: If they had a distinctive value system satisfying, say, four of the listed items, and for this system there was a word in their language—say "woogle values"—would we translate "woogle" into "moral"? It's not my place to guess with any confidence how that counterfactual decision would go. All I am claiming is that the foregoing items would all be important considerations in that decision.

What evidence is there that the human proclivity for making such judgments is innate? The reader could be forgiven for assuming that an examination of such empirical evidence will be the focus of this chapter, but in fact this is another matter concerning which I must content myself with a wave of the hand in a certain direction. On this occasion my objective is not to attempt to establish that human morality *is* innate, but rather to address the question of how and why it *could* be: What makes moral judgment adaptive, and what evolutionary forces might have been involved in its emergence? Having a good answer to these questions does in itself provide some support for the hypothesis that morality is innate, for this hypothesis would be shaky if we lacked any conception of how natural selection might have produced such a trait. Nevertheless, of course, having a coherent story to tell about how a trait *could have* resulted from natural selection is never sufficient for establishing that it did so evolve. For that we need hard evidence. In my opinion (here comes the hand-waving), the strongest evidence for an innate human faculty comes from developmental psychology. The course of moral development in the human child exhibits an extremely reliable sequence, it gets underway remarkably early, its developmental pathway is distinct from the emergence of other skills, and its unfolding includes abrupt maturations. On this last point, Jonathan Haidt (2001, p. 826-827) describes the view of anthropologist Alan Fiske (1991) as follows:

> …children seem relatively insensitive to issues of fairness until around the age of 4, at which point concerns about fairness burst forth and are overgeneralized to social situations in which they were never encouraged and in which they are often inappropriate. This pattern of sudden similarly timed emergence with overgeneralization suggests the maturation of an endogenous ability rather than the learning of a set of cultural norms.

Of particular note is the child's capacity to distinguish moral from conventional transgressions, which emerges as early as the third year (Smetana, 1981; Smetana and Braeges, 1990)—and this is an impressively cross-cultural phenomenon (Nucci et al., 1983; Hollos et al., 1986; Song et al., 1987; Yau and Smetana, 2003). Whence do children derive this distinction? It is exceedingly unlikely that across the wide variety of human social ecologies there is some stable exogenous characteristic that may be plausibly appealed to as the explanans of this developmental phenomenon. For example, one of the features taken to distinguish the moral from the conventional is the independence of moral normativity from any rule-conferring authority figure (see Turiel, 1983, 1998; Turiel et al., 1987). Yet it is difficult to see what there might be in a typical social environment that would allow a "general intelligence mechanism" to infer on the basis of observation that one norm depends on authoritative decree (e.g., that boys should not wear dresses to school) while another does not (e.g., that one shouldn't punch others). In order to infer a *dependence* relation, one would have to observe a correlation between the relevant authority's changing its mind to permit the boy to wear a dress and that action's no longer counting as a transgression. And in order to infer an *independence* relation one would have to either (1) observe the relevant authority change its opinion about an act of harming while one noted that the act nevertheless continued to count as a transgression, or (2) observe a previously-condemned act of harming cease to count as a transgression (or vice versa) while one noted that the relevant authority's opinion on the matter had not altered. But observations of types 1 and 2 are hard to come by, even for adults, let alone three-year-olds. Regarding a serious moral offense, like violent crime, what we invariably observe is both elements remaining stable: All relevant authorities denounce it, and it continues to be considered a transgression. How, on the basis of such observations, a child is supposed to infer an independence relation is baffling.[3] The solution to this puzzle is that morality is not something that children learn or infer from their exogenous environment but is, rather, the result of the unfolding of an innate preparedness.

As I say, rather than develop this line of argument (or any of a number of complementary lines of argument), what I intend in this chapter is to ask why natural selection might have been interested in producing such a trait. A group selectionist account will be satisfactory as an explanation if it shows how having individuals making such authoritative prosocial judgments would serve the interests of the group. An explanation in terms of individual selection must show how wielding authoritative prosocial judgments would enhance the inclusive reproductive fitness of the individual. One might be tempted to think that the group selectionist account is more feasible since it can more smoothly explain the development of prosocial instincts—after all, it is virtually a tautology that prosocial tendencies will serve the interests of the group. However, prosociality may also be smoothly explained in terms of individual selection via an appeal to the processes of kin selection, mutualism, and reciprocal altruism (see Dugatkin, 1999). In what follows I will focus on the last.

---

[3] Likewise, what experience allows a child to infer that certain norms are local whereas others hold more generally (this being another criterion for distinguishing conventional norms from moral)? When the locale of the norm is, for example, school versus home, we can plausibly find the origin of the distinction in the child's experience. But many social conventions hold in both the school and the home, and in fact for a wide range of social norms (e.g., eating with utensils rather than fingers), the child very often has neither direct nor indirect experience of a setting in which it doesn't hold.

## 2 Reciprocity

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

It is often thought that cheating and "cheat-detection" traits are an inevitable or even defining feature of reciprocal exchanges, but in fact a relationship whose cost-benefit structure is that of reciprocal altruism could in principle exist between plants—organisms with no capacity to cheat, thus prompting no selective pressure in favor of a capacity to detect cheats. Even with creatures who have the cognitive plasticity to cheat on occasions, reciprocal relations need not be vulnerable to exploitation. If the cost of cheating is the forfeiture of a highly beneficial exchange relation, then any pressure in favor of cheating is easily outweighed by a competing pressure against cheating, and if this is reliably so for both partners in an ongoing program of exchange, then natural selection doesn't have to bother giving either interactant the temptation to cheat, or a heuristic for responding to cheats. But since reciprocal exchanges will develop only if the costs and benefits are balanced along several scales, and since values are rarely stable in the real world, there is often the possibility that a reciprocal relation will collapse if environmental factors shift. If one partner, A, indicates that he will help others no matter what, then it may no longer be to B's advantage to help A back. If the value of cheating were to rise (say, if B could possibly *eat* A, and there's suddenly a serious food shortage), then it may no longer be to B's advantage to help A back. If the cost of seeking out new partners who would offer help (albeit only until they also are cheated) were negligible, then it may no longer be to B's advantage to help A back. For natural selection to favor the development of an ongoing exchange relation, these values must remain stable and symmetrical for both interactants.[4] What is interesting about

[4] By "symmetrical" I mean that it is true of each party that she is receiving more benefit than cost incurred. But it is in principle possible that, all told, one of the interactants is getting vastly more benefit than the other. Suppose B gives A 4 units of help, and it costs him 100 units to do so. Sounds like a rotten deal? Not if we also suppose that A in return gives B 150 units of help, and it costs her only 3 units to do so. Despite the apparent unevenness of the exchange, since 4 > 3 and 150 > 100, both players are up on the deal, and, *ceteris paribus*, they should continue with the arrangement. The common assumption—that what is vital to reciprocal exchanges is that one can give a benefit for relatively little cost—need not be true of *both* interactants. With the values just given, it is not true of B. But when it is not true of one of the interactants, then in order to compensate it must be "very true" of the other: Here A gives 150 units for the cost of only 3.

many reciprocal arrangements is that there's a genuine possibility that one partner can cheat on the deal (once she has received her benefit) and get away with it. Therefore there will often be a selective pressure in favor of developing a capacity for distinguishing between cheating that leads to long-term forfeiture and cheating that promises to pay off. This in turn creates a new pressure for a sensitivity to cheats and a capacity to respond to them. An exchange between creatures bearing such capacities is a *calculated* reciprocal relationship; the individual interactants have the capacity to tailor their responses to perceived shifts in the cost-benefit structure of the exchange (see de Waal and Luttrell, 1988).

The cost-benefit structure of a reciprocal relation can be stabilized if the price of non-reciprocation is increased beyond the loss of an ongoing exchange relationship. One possibility would be if individuals actively punished anyone they have helped but who has not offered help in return. Another way would be to punish (or refuse to help[5]) any individual in whom you have observed a "non-reciprocating" trait, even if you haven't personally been exploited. One might go even further, punishing anyone who refuses to punish such non-helpers. The development of such punishing traits may be hindered by the possibility of "higher order defection," since the individual who reciprocates but doesn't take the trouble to punish non-reciprocators will apparently have a higher fitness than reciprocators who also administer the punishments. Robert Boyd and Peter Richerson (1992) have shown that this is not a problem so long as the group is small enough that the negative consequences of letting non-reciprocators go unpunished will be sufficiently felt by all group members. They argue, however, that we must appeal to cultural group selection in order to explain punishing traits in larger groups. I have two things to say in response to this last point. First, the reason that increased group size has such an impact on the effectiveness of punishment strategies is that the multiplication of interactants amplifies the costs of coercion. But if an increase in group size is accompanied by the evolution of a trait that allows an individual to spread her punishments more widely at no extra cost, then this consideration is mitigated. It has been argued (with much plausibility, in my opinion) that language is precisely such a mechanism (see Aiello and Dunbar, 1993; Dunbar, 1993, 1996; Smith, 2003). Talk, as they say, is cheap, but it allows one to do great harm to the reputation of a virtual stranger. Second, on the assumption that through the relevant period of genetic natural selection our ancestors lived in relatively small bands—small enough, at least, that a person not pulling his or her weight was a burden on the group—Boyd and Richerson's cogent argument doesn't undermine the hypothesis that an innate human morality can be explained by reference only to individual selection. Perhaps they are correct that cultural group selection must be invoked to explain the explosion of human ultra-sociality in the Holocene; and perhaps it is a process that has contributed a great deal to the content of moral codes. But neither observation is at odds with

[5] In some scenarios there may not be much difference in refusing help and punishing, despite one sounding more "active" than the other. If a group of, say, baboons were to terminate all interactions with one of their troop, this would penalize the ostracized individual as much as if they killed the individual outright. This is one reason why I am troubled by Chandra Sripada's efforts to place reciprocity-based and punishment-based accounts of moral compliance *in opposition* to each other (2005). Punishment will often be a natural concomitant of reciprocity—as even Trivers noted in his 1971 paper. It should also be noted that "refusing to play" can be as costly as administering punishment. If lions were to refuse to share with a free-riding lioness, then they would have to drive her off when she barged in to share their kill, perhaps risking injury to do so. (As a matter of fact, it turns out that lions are rather tolerant of free-riders; their helping behaviors seem regulated by mutualism rather than reciprocation. See Heinsohn and Packer 1995.)

my hypothesis, since it may be maintained that a biological human moral sense antedates the large-scale ultra-sociality of modern humans. Indeed, Boyd and Richerson as much as admit this when they allow that "moral emotions like shame and a capacity to learn and internalize local practices" existed as genetically coded traits prior to any spectacular cultural evolution (Richerson et al., 2003, p. 371).

Another trait that might be expected to develop in creatures designed for reciprocation is a faculty dedicated to the acquisition of relevant information about prospective exchange partners prior to committing to a relationship. Gathering social information may cost something (in fitness terms), but the rewards of having advance warning about what kind of strategy your partner is likely to deploy may be considerable. This lies at the heart of Richard Alexander's account (1987) of the evolution of moral systems. In *indirect* reciprocal exchanges, an organism benefits from helping another by being paid back a benefit of greater value than the cost of her initial helping, but not necessarily by the recipient of the help. We can see that reputations involve indirect reciprocity by considering the following: Suppose A acts generously towards several conspecifics, and this is observed or heard about by C. C, meanwhile, also learns of B's acting disreputably towards others. On the basis of these observations—on the basis, that is, of A's and B's reputations—C chooses A over B as a partner in a mutually beneficial exchange relationship. A's costly helpfulness has thus been rewarded with concrete benefits, but not by those individuals to whom he was helpful. Alexander lists three major forms of indirect reciprocity:

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

One possible example of indirect reciprocity is the behavior of Arabian babblers, as studied by Amotz Zahavi over many years (Zahavi and Zahavi, 1997). Babblers are social birds that act in helpful ways towards each other: feeding others, acting as sentinels, etc. What struck Zahavi was not this helpful behavior *per se*, but the fact that certain babblers seem positively eager to help: jostling to act as sentinel, thrusting food upon unwilling recipients. The "Handicap Principle" that Zahavi developed states that such individuals are attempting to raise their own prestige within the group: signaling "Look at me; I'm so strong and confident that I can afford such extravagant sacrifices!" Such displays of robust health are likely to attract the attention of potential mates while deterring rivals, and thus such behavior is, appearances notwithstanding, squarely in the fitness-advancing camp.[6]

---

[6] The connection between indirect reciprocity and the Handicap Principle is commented on by Nowak and Sigmund, 1998.

Consider the enormous and cumbersome affair that is the peacock's tail. Its existence poses a *prima facie* threat to the theory of natural selection—so much so that Charles Darwin once admitted that the sight of a feather from a peacock's tail "makes me sick!" (F. Darwin 1887, p. 296). Yet Darwin also largely solved the problem by realizing that the primary selective force involved in the development of the peacock's tail is the peahen's choosiness in picking a mate.[7] If peahens prefer mates with big fan-shaped tails, then eventually peacocks will have big fan-shaped tails; if peahens prefer mates with triple-crested, spiraling, red, white, and blue tails, then (*ceteris paribus*) eventually peacocks will sport just such tails. Sexual selection is a process whereby the choosiness of mates or the competition among rivals can produce traits that would otherwise be detrimental to their bearer. I am not categorizing sexual selection in general as reciprocity, only those examples that involve the favoring of traits of costly helpfulness. If a male is helpful to a female (bringing her food, etc.) and, as a result, she confers on him the proportionally greater benefit of reproduction, this is an example of direct reciprocity. If a male is helpful to his fellows in general and, as a result, an observant female confers on him the proportionally greater benefit of reproduction (thus producing sons who are generally helpful and daughters who have a preference for helpful males), this is an example of indirect reciprocity. Just as sexual selection can produce extremely cumbersome physical traits, like the peacock's tail, so too can it produce extremely costly helping behaviors. We can say the same of reputation in general if the benefits of a good reputation are great enough. If a good reputation means sharing food indiscriminately with the group, then an indiscriminate food-sharing trait will develop; if a good reputation means wearing a pumpkin on your head, then a pumpkin-wearing trait will develop. The same, moreover, can be said of punishment, which is, after all, the flip side of being rewarded for a good reputation. If a type of self-advancing behavior (or any type of behavior at all) is sufficiently punished, it will no longer be self-advancing at all (see Boyd and Richerson, 1992).

Once we see that indirect reciprocity encompasses systems involving reputation and punishment, and that these pressures can lead to the development of just about *any* trait—extremely costly indiscriminate helpfulness included—then we recognize what a potentially vital explanatory framework it is. It is important to note, however, that all that has been provided in this section is an account of a process whereby prosocial behavior can evolve; the organisms designed to participate in such relations might be insects—they need not have a *moral* thought in their heads.

## 3 Reciprocity and altruism

The view I am interested in advocating is that in cognitively advanced creatures moral judgment may add something to reciprocal exchanges: It may contribute to their success in a fitness-enhancing manner, such that a creature for whom reciprocal relations are important may do better with a sense of *obligation* and *prohibition* guiding her exchanges than she would if

[7] I say "*largely* solved" since Darwin did not present an explanation of why it is the *female* who gets to be the choosy one. The answer is that in many species females must invest a lot of energy in their offspring, whereas males can hope to get away with investing very little. This answer was, I believe, first appreciated by the early geneticist Ronald Fisher ([1930] 1999).

motivated solely by "unmoralized" preferences and emotions. The advantages of reciprocity, then, may have provided the principal selective pressure that produced the human moral sense.

Before proceeding, however, a couple of quick objections to the hypothesis should be nipped in the bud. First, it might be protested that many present-day moral practices have little to do with reciprocation: Our duties to children, to the severely disabled, to future generations, to animals, and (if you like) to the environment all are arguably maintained without expectation of payback. Yet this objection really misses the mark, for these considerations hardly undermine the hypothesis that it was for regulating reciprocal exchanges that morality evolved in the first place; it is not being claimed that reciprocity alone is what continues to sustain social relations. Reciprocity may give someone a sense of duty towards his fellows that causes him to hurl himself on a grenade to save their lives. There is no actual act of reciprocation there—not even an expectation of one—but nevertheless reciprocity may be the process that brought about the psychological mechanisms that prompted the sacrificial behavior. Although these mechanisms may have evolved in order to govern reciprocal exchanges (producing, we might expect, judgments that are highly dependent on what kind of relation the individuals stand in), it should come as no surprise that social factors might develop that urge, say, a more universal benevolent attitude—perhaps even encouraging one to initiate and continue relations irrespective of one's partner's actions (e.g., to turn the other cheek). By comparison, one might hypothesize that human color vision evolved in order to allow us to distinguish ripe from unripe fruit, but this would hardly imply that this continues to be the only thing we can do with color vision.

Second, it might be objected that a person enters into a reciprocal relationship for self-gain, and thus is motivated entirely by selfish ends (albeit perhaps "enlightened self-interest")—the very antithesis of *moral* thinking. This objection is confused. Entering into reciprocal relations may well be fitness-advancing, but this implies nothing about the motivations of individuals designed to participate in such relations. Even Darwin got this one wrong: In the passage from *The Descent of Man* often cited as evidence of his appreciation of the importance of reciprocity in human prehistory, he attributes its origins to a "low motive" (Darwin [1879] 2004, p. 156).[8] George Williams (1966, p. 94) correctly responds: "I see no reason why a conscious motive need be involved. It is necessary that help provided to others be occasionally reciprocated if it is to be favored by natural selection. It is not necessary that either the giver or the receiver be aware of this." I would add that I see no reason that an *unconscious* motive need be involved either. In vernacular English, whether an action is "selfish" or "altruistic" depends largely (if not entirely) on the motives with which it is performed. (Suppose Amy acts in a way that benefits Bert, but what prompts the action is her belief that she will benefit herself in the long run. Then it is not an altruistic act, but a selfish act. Suppose Amy's belief turns out to be false, so that she never receives the pay-off and the only person who gains from her action is Bert. This does not cause us to retract the judgment that her action was selfish.) It follows that creatures whose cognitive lives are sufficiently crude that they lack such deliberative motives cannot be selfish or altruistic in this everyday sense at all, and yet they may very well be involved in reciprocal exchanges.

[8] This perhaps should be put down to a sloppy choice of wording, for elsewhere in *Descent* Darwin argues staunchly against psychological egoism.

It is standard to distinguish altruism in this psychological sense from "evolutionary altruism," which is an altogether more complex and controversial affair, consisting of a creature lowering its inclusive reproductive fitness while enhancing the fitness of another.[9] Reciprocal altruism is not an example of evolutionary altruism (see Sober, 1988); in a reciprocal exchange, neither party forfeits fitness for the sake of another. As Trivers defined it, "altruistic behavior" (by which he means *helpful* behavior) is that which is "apparently detrimental to the organism performing the behavior" (1971, p. 35)—but obviously an *apparent* fitness-sacrifice is not an actual fitness-sacrifice, any more than an apparent Rolex is an actual Rolex. Others have defined "reciprocal altruism" as fitness-sacrificing *in the short term*. But again: Foregoing a short-term value in the expectation of greater long term gains is no more an instance of a genuine fitness sacrifice than is, say, a monkey's taking the effort to climb a tree in the hope of finding fruit at the top. So despite claims that reciprocal altruism and kin selection together solve the so-called *paradox of evolutionary altruism*, if (i) by "altruism" we mean *fitness sacrificing* (not *apparent* or *short-term* fitness sacrificing), and (ii) by "fitness" we mean inclusive fitness, and (iii) by "*solving* the paradox of evolutionary altruism" we mean showing how such altruism is possible, then I see no reason at all for thinking that this frequently repeated claim is true.

But if reciprocal altruism is altruism in neither the vernacular nor the evolutionary sense, then in what sense is it altruism at all? The answer is that it is not. I have called it "reciprocal altruism" in deference to a tradition of 30 years, but in fact I don't like the term, and much prefer to call it "reciprocal exchanges" or just "reciprocity." What it is is a process by which *cooperative* and *helpful* behaviors evolve, not (necessarily) a process by which altruism evolves. I add the parenthetical "necessarily" because it *may* be that in cognitively sophisticated creatures, altruism, in the vernacular sense, may evolve as a proximate mechanism for regulating such relations, but it is certainly no necessary part of the process, since it is also possible that for some intelligent creatures the most efficient way of running a reciprocal exchange program is to be deliberatively Machiavellian—i.e., selfish in the vernacular sense. My point is that neither motivational structure can be inferred from the fact that a creature is designed to participate in reciprocal exchanges. Reciprocal partners may enter into such exchanges for selfish motives, or for altruistic motives, or their exchanges may be mere conditioned or hard-wired reflexes properly described neither as selfish nor altruistic. Genes inhabiting selfishly-motivated reciprocating organisms may be soundly out-competed by genes inhabiting reciprocating organisms who are moved directly by the welfare of their chosen exchange partners. And genes inhabiting reciprocating organisms motivated additionally by thoughts of moral duty, who will feel guilty if they defect, may do better still.

## 4 Ancestral reciprocity

The lives of our ancestors over the past few million years display many characteristics favorable to the development of reciprocity. They lived in small bands, meaning that they would interact

[9] On the face of it, evolutionary altruism, as it is here defined, seems impossible. Sober and Wilson (1999) argue that it is possible only by invoking group selection, and so long as we take care to avoid what they call "the averaging fallacy" (1999, p. 31-35). Even if their argument is successful, however, it remains an open question how much of the prosocial behavior observable in nature (bees, ants, humans, etc.)—which is often casually referred to as "altruism"—is an instance of evolutionary altruism.

with the same individuals repeatedly. The range of potential new interactants was very limited, thus the option of cheating one's partner in the expectation of finding another with whom one could enter into exchanges (perhaps also to cheat) was curtailed. We can assume that interactions were on the whole quite public, so opportunities for secret uncooperative behaviors were limited. They lived relatively long lives—long enough, at least, that histories of interaction could develop—and they probably had relatively good memories. Some of the important foods they were exploiting came unpredictably in large "packages"—i.e., big dead animals—meaning that one individual, or group of individuals, would have a great deal of food available at a time when others did not, but in all likelihood at a later date the situation would be reversed. Large predators were a problem, and shared vigilance and defense was a natural solution. Infants required a great deal of care, and youngsters a lot of instruction. Though we don't need to appeal to reciprocity to explain food sharing, predation defense, or childrearing, what these observations do imply is that there were available several basic forms of "currency" in which favors could be bestowed and repaid. This means that someone who was, say, unable to hunt could nevertheless repay the services of the hunter in some other form. If we factor in the development of language, then we can add another basic currency: the value of shared information. All these kinds of exchanges (the last in particular) allow for the "give-a-large-benefit-for-a-relatively-low-cost" pattern that is needed for reciprocity to be viable.

When we start to list such characteristics, what emerges is a picture of an animal ripe for the development of reciprocity—indeed, it is hard to imagine any other animal for whom the conditions are so suitable. Bearing in mind the enormous potential of reciprocity to enhance fitness, we might suspect natural selection to have taken an interest, to have endowed our ancestors (and thus us) with the psychological skills necessary to engage efficiently in such relations. What kind of skills might these be? We have already mentioned some: a tendency to look for cheating possibilities; a sensitivity to cheats, a capacity to remember them, and an antipathy towards them; an interest in acquiring knowledge of others' reputations, and of broadcasting one's own good reputation. We can add to these a sense of distributive fairness; the capacity to distinguish accidental from intentional "defections" and an inclination to forgive injuries of the former kind; and if those participating in a reciprocal exchange are trading concrete goods, then we would expect a heightened sense of ownership to develop.

Here is not the place to review empirical evidence favoring the view that the human mind has evolved such tendencies; such support comes from a number of fields: developmental psychology, neuroscience, cross-cultural anthropology, experimental economics, evolutionary psychology, primatology. Let me, however, very briefly gesture towards some evidence pertaining to the last item mentioned—a sense of ownership—on the grounds that the role of this trait in the evolution of human reciprocity seems under-appreciated in the literature, as, indeed, does the fact that *ownership* (as opposed to mere *possession*) is a highly moralized relation. To the extent that trade implies a grasp of ownership, we find the physical traces of ownership far back in the archaeological record, at least into the early Upper Paleolithic (Mellars, 1995, p. 398-400), and perhaps far beyond (McBrearty and Brooks, 2001). There is not a shred of evidence that trade (or reciprocity more generally) is a *de novo* artifact of modern civilization that spread from one or more points of cultural invention. It is, rather, like language: ubiquitous and

ancient.[10] A sense of ownership, moreover, emerges more or less spontaneously in the course of childhood development, and surprisingly early: the very first two-word linguistic strings that an infant manages to construct and comprehend often denote ownership relations (e.g., "Mommy sock" for *Mommy's sock*) (see Brown, 1973; Markessini and Golinkoff, 1980). Numerous studies have shown that the vast majority of playroom conflicts among children concern possession of items, beginning as early as the children are capable of generating any kind of interpersonal conflict at all (see Dawe, 1934; Bronson, 1975; Smith and Green, 1975). The few grand social experiments that have attempted to expunge the notion of ownership from the human psyche—such as in the Soviet Union or the kibbutzim of Israel—have encountered an extremely stubborn opponent. Discussing this phenomenon in the 1950s, the anthropologist Melford Spiro wrote:

> the child is no *tabula rasa*, who, depending on his cultural environment, is equally amenable to private or collective property arrangements. On the contrary, the data suggest that the child's early motivations are strongly directed towards private ownership, an orientation from which he is only gradually weaned by effective cultural techniques.
>
> (Spiro, 1958, p. 375-6)

In admitting that this amounts to no more than a gesture toward the kind of evidence we should be looking for, I don't mean to suggest that there is a large and overwhelming body of evidence that I'm skirting in the interests of brevity. Whether there really are parts of the human mind dedicated to ownership or reciprocal exchanges in general, or whether such universal skills are instead the product of our general all-purpose intelligence, remains to be established, and doing so will not be easy. What we should not expect from anyone is a deductive argument from demonstrably true premises; rather, we should hope for a "picture" of the human mind that fits well with the available evidence and promises to help us make sense of things. But at least one thing is clear: There is enough evidence supporting this hypothesis that the tired sneer that it is merely a "Just So Story" is no longer warranted. It is a plausible, coherent, productive, and testable hypothesis, and there is good reason for looking favorably upon it.

## 5 Morality and motivation

But what's morality got to do with it? What is added to the stability of a reciprocal exchange if the interactants think of cheating as "morally odious" (say), as opposed to them simply having a strong "unmoralized" disinclination to cheat? Note that this is a pressing question not just for the advocate of the hypothesis presently under discussion, but is a good question for *anyone*, even those who think that morality is a purely cultural construct. What practical benefit does distinctively moral thinking bring? Someone seeking to explain morality as a biological

[10] Sometimes we hear tell of societies with no sense of private ownership, but upon examination it turns out that these societies just own different things than we (in the West) are familiar with. Certainly there are cultures where *land* isn't an owned item, and cultures where there are very few possessions, but there is no human society where the very idea of an item being owned (be it only articles of clothing, weapons, or a few ornaments) is unknown. Other cultures may also more readily employ the concept of *collective* ownership—but, of course, goods belonging to the family or the tribe are just as much conceived of as property as those belonging to an individual. As a matter of fact, however, the concept of *individual* ownership appears to be a human universal.

phenomenon and invoking only individual selection may find it useful to tease apart two questions: What benefit does an individual gain by judging *others* in moral terms? What benefit does an individual gain by judging *himself* in moral terms? I will start out addressing the latter question, though the need to tie this to a discussion of the former will quickly become apparent.

It is natural to suppose that an individual's sincerely judging some available action in a morally positive light increases her probability of performing that action (likewise, *mutatis mutandis*, judging an action in a morally negative light). If reproductive fitness will be served by the performance or the omission of a certain action, then it will be served by any psychological mechanism that ensures or probabilifies this performance or omission (relative to mechanisms that do so less effectively). Thus self-directed moral judgment may enhance reproductive fitness so long as it is attached to the appropriate actions. We have already seen that the "appropriate actions"—that is, the fitness enhancing actions—will in many circumstances include helpful and cooperative behaviors. Therefore it may serve an individual's fitness to judge certain prosocial behaviors—*her own* prosocial behaviors—in moral terms.

The part of the foregoing case that needs development is the premise that moral judgment probabilifies the performance or omission of actions. There is plenty of empirical evidence to this effect (see Keltner et al., 1995; Bandura et al., 1996; Bandura, 1999; Ferguson et al., 1999; Tangney, 2001; Beer et al., 2003; Keltner, 2003; Covert et al., 2003; Ketelaar and Au, 2003), but in what follows I will develop the argument along a particular avenue.

The benefits that may come from cooperation—enhanced reputation, for example—are typically long-term values, and merely to be aware of and desire these long-term advantages does not guarantee that the goal will be effectively pursued, any more than the firm desire to live a long life guarantees that a person will give up fatty foods. (The human tendency to discount future gains is well-documented: see Schelling, 1980; Elster, 1984; Ainslie, 1992.) Self-directed moral judgment often does better than long-term prudential deliberation in securing the correct motivations. If you are thinking of an outcome in terms of something that you desire, you can always say to yourself "But maybe foregoing the satisfaction of that desire wouldn't be *that* terrible." If, however, you're thinking of the outcome as something that is *desirable*—as having the quality of *demanding* desire—then your scope for rationalizing a spur-of-the-moment devaluation narrows. When a person believes that an act of cooperation is *morally* required—that it *must* be practiced whether he likes it or not—then the possibilities for further internal negotiation on the matter diminish. If a person believes an action to be required by an authority from which he cannot escape, if he believes that in not performing it he will not merely frustrate himself, but will become *reprehensible* and *deserving of disapprobation*—then he is more likely to perform the action. The distinctive value of imperatives imbued with such practical clout is that they silence further calculation, which is a valuable thing when our prudential calculations can so easily be hijacked by interfering forces and rationalizations. What is being suggested, then, is that self-directed moral judgments can act as a kind of personal commitment, in that thinking of one's actions in moral terms eliminates certain practical possibilities from the space of deliberative reasoning in a way that thinking "I just don't like X" does not.[11] In saying this I

[11] Note that the argument doesn't depend on comparing someone who is motivated by non-moralized sympathy with someone who is utterly unsympathetic but has a robust rational sense of moral duty—a thought experiment familiar to students of Kant. First, we are granting the moralized person all the sympathies and inclinations of the

am in part agreeing with Daniel Dennett (1995), who argues that moral principles function as "conversation-stoppers": considerations that can be dropped into a decision process (be it a personal or interpersonal decision) in order to stop mechanisms or people from endlessly processing, endlessly reconsidering, endlessly asking for further justification. "Any policy *may* be questioned, so, unless we provide for some brute and a-rational termination of the issue, we will design a decision process that spirals fruitlessly to infinity" (Dennett, 1995, p. 506). In deciding how to treat a criminal, the consideration "He has a moral right to a fair trial" seems to close off further discussion. In deciding whether to shoplift, the consideration "It is wrong to shoplift; I mustn't do it" puts an end to deliberations. "Faced with a world in which such predicaments are not unknown," says Dennett, "we can recognize the appeal of … some unquestioning dogmatism that will render agents impervious to the subtle invasions of hyper-rationality" (1995, p. 508).

These thoughts, however, provide only half the answer to the question we are addressing, for one might still wonder what it is about a moral judgment that makes it function so well as a conversation-stopper. Presumably *non-moral* considerations also often function effectively in this manner; the thought "I would die if I did that" will in most circumstances put an end to any further deliberations in favor of performing the action in question. One way of putting this worry is to ask what motivation-strengthening features moral judgment has that strong (but non-moral) desire does not have. The worry deepens when we bear in mind that nothing I have said is intended to undermine the truism that what ultimately determines whether a person acts is the strength of her desires in favor of so acting compared with her desires against acting; the hypothesis being advocated is that moral judgment bolsters desire. This, then, leaves us with the question—posed by David Lahti (2003)—of why natural selection did not simply make humans with stronger desires that directly favor cooperation in certain circumstances. After all, for some adaptive behaviors this is precisely what evolution has granted us. Protective actions towards our offspring, for example, appear to be regulated by robust raw emotions, not primarily by any moralistic sense of duty. These emotions are by and large stoutly resistant to the lures of weakness of will: Few are tempted to rationalize a course of action that promises short-term gain while resulting in injury to their beloved infant. Moreover, insofar as our hominid forebears already had in place the neurological mechanisms for such strong desires, it's something of a mystery why the inherently conservative force of natural selection would not press into service these extant mechanisms in order to govern any novel adaptive behavior, rather than fabricating a "radically different" and "biologically unprecedented mechanism for a purpose which is achieved regularly in nature by much more straightforward means" (Lahti, 2003, p. 644). Lahti's challenge must be addressed.

---

non-moralized person; the argument is just that moral judgment *adds something* to that motivational profile, that it gives her an edge. Nor is the claim that moral thinking always does better than prudential thinking, for a lot of the time prudential thinking is completely resolute (the knowledge that crossing the highway will result in your death is probably more motivationally engaging than the judgment that jaywalking is morally forbidden); the argument is just that moral judgment can step in on those occasions when prudence may falter (in particular when the prudential gain is a probabilistic long-term affair). Also it must be remembered that moral judgment is not being conceived of here as the cool intellectualized affair that Kant fancied it to be; an element of what self-directed moral judgment adds to a person's mental life, for example, is the emotion of guilt. When I say that moral judgment promotes motivation, I am including the motivational efficacy of certain moral emotions.

Whenever an evolutionary psychologist hypothesizes about the presence of a specialized mechanism functioning to govern an adaptive behavior, the query can always be raised: "Why would natural selection bother with that mechanism? Why wouldn't it simply create an overwhelmingly strong desire to perform that behavior?" That there is something fishy about this question is revealed if we consider some non-moral cases. Think instead about the psychological reward systems that have evolved in humans regarding sex and eating. One might ask why natural selection bothered giving us all that complicated physiological equipment needed for having an orgasm—why not design us simply to *want* to have sex? It seems a misguided question. Natural selection *did* make us want to have sex, and one of its means of ensuring this desire was precisely the human orgasm. Similarly, natural selection made us want to eat food, and one of its means of achieving this was to create a creature for whom food tastes good and hunger feels bad. And perhaps natural selection has made us want to cooperate, and granting us a tendency to think of cooperation in moral terms is a means of securing this desire. That natural selection may employ a distinctive *means* for creating and strengthening a type of fitness-advancing desire is no more mysterious in the moral case than in the other two cases. Granted, in the moral case we are considering a "biologically unprecedented mechanism"—something that evolved uniquely in the hominid line—but insofar as human social relations *are* radically different from those of other animals, a radically different solution may have been necessary. Note also that despite the conservatism of natural selection, there is an obvious reason that distinct fitness-advancing behaviors will often require different mechanisms motivating them: If eating or promise-keeping were rewarded with an orgasm, then an individual might not bother with sex.

It is still reasonable to inquire what special features a moral judgment might have that render it suited to the evolutionary task we are speculatively assigning it here. An important part of the answer, I think, concerns the public nature of moral judgments. That we are now focusing on self-directed moral judgments shouldn't lead us to assume that we are talking about a private mental phenomenon. There can be private other-directed judgments (e.g., ruminating quietly to oneself "John's such a bastard"), just as there can be publicly announced self-directed judgments ("I want you all to know that I'm thoroughly ashamed of what I did"). A moral judgment, even a self-directed one, is essentially communicative; it is something that may be asserted in the course of collective negotiation, may be employed to stake a claim, to justify a decision, to provide warrant for a punishment, to criticize or praise another's conduct or character, or to present evidence of one's own character. The manner in which thinking of a possible course of action in morally positive terms promotes the motivation to perform it cannot be divorced from this public sphere. Even when my private conscience guides me to refrain from cheating with the thought "Cheating is wrong," I am aware that this is a consideration that might be brought into the domain of public deliberation if I am required to justify my actions; I am accepting that, were I to cheat, punishment from others would be warranted. By comparison, the proposition "I just don't like cheating" may be brought forward to *explain* one's actions, but it lacks the normative *justificatory* force of a moral consideration.[12] A person's resolve to act (or not to act) is

[12] "I really don't like X" can be *an element of* a justification: "I really don't like X, and in these circumstances it is acceptable for my actions to be guided by my strong preferences." Clearly, though, the latter part of the justification introduces a normative principle. Often the latter part will be tacit: "I like coffee" can seem like a

importantly affected by her conception of how others will receive her decisions, her confidence in whom she can justify herself to, her perception of herself as acting from considerations that would also move her fellows—in short, her experience of herself as a social being. Lahti's puzzle is solved when we realize that a moral judgment affects motivation not by giving an extra little private mental nudge in favor of certain courses of action, but by providing a deliberative consideration that (putatively) cannot be legitimately ignored, thus allowing moral judgments—even self-directed ones—to play a justificatory role on a social stage in a way that unmediated desires cannot.

This reasoning leads me to supplement the simple hypothesis with which we started: that the evolutionary function of moral judgment is to provide added motivation in favor of certain adaptive social behaviors. Morally disapproving of one's own action (or potential action)—as opposed to disliking that action—provides a basis for corresponding *other*-directed moral judgments. No matter how much I dislike something, this inclination alone is not relevant to my judgments concerning *others* pursuing that thing: "I won't pursue X because I don't like X" makes perfect sense, but "You won't pursue X because I don't like X" makes little sense. By comparison, the assertion of "The pursuit of X is morally wrong" demands both *my* avoidance of X *and yours*. By providing a framework within which both one's own and others' actions may be evaluated, moral judgments can act as a kind of "common currency" for collective negotiation and decision-making. Moral judgment thus can function as a kind of social glue: bonding individuals together in a shared justificatory structure, providing a tool for solving many group coordination problems. Of particular importance is the fact that although a non-moralized strong negative emotional reaction (e.g., anger) may prompt a punitive response, it takes a moral judgment to supply *license* for punishment, and thus the latter serves far more effectively to govern public decisions in a large group than do non-moralized emotions or desires.

One final thing that should be emphasized is that although for brevity's sake I have spoken of moral judgments as bolstering the motivation to cooperate, I don't mean to imply that we are designed to be *unconditional* cooperators. The moral sense is not a proclivity to judge cooperation as morally good in any circumstance—something that looks like a recipe for disastrous exploitation. By the same token, the fact that we have innate mechanisms dedicated to making us want to eat, rewarding us with pleasure for doing so, doesn't mean that we eat unconditionally and indiscriminately. We may be designed to be very plastic with respect to cooperative strategies. How generous one can afford to be, or how miserly one is forced to be, will depend on how resource-rich is one's environment. Who is a promising partner and who is a scoundrel is something we learn. One can moralize a conditional strategy, such as "Be trusting, but don't be a sucker." One can moralize non-cooperation, seeing it as forbidden in certain circumstances. The idea being advocated is that there are adaptive benefits to be had by moralizing the whole plastic social structure. Doing so prevents under-performance, which is not to be confused with encouraging over-performance. It is true that there is a sense in which any boost to the motive to cooperate on a token occasion means that one may be encouraged to commit a practical error—to stick with an exchange relation when one's fitness would really be

---

perfectly good justification alone for drinking coffee, but that there is an unspoken premise here (to the effect that one is in circumstances where preferences may legitimately guide action) is obvious if we compare "I like torturing children."

better served by cheating. But this is the same sense in which any natural reward system can lead us to occasional and even disastrous error: The craving for food can lead someone to eat a poisonous plant, and the pleasures of sex can result in making powerful enemies.

## 6 Group selection

I should like to end by commenting on the comparison between the hypothesis outlined in this chapter—a hypothesis ostensibly in terms of individual selection—and the well-known views on group selection put forward by Elliott Sober and David Sloan Wilson in *Unto Others* (1999). I will confine myself to three points.

1) Sober and Wilson do not purport to put forward a theory concerning the evolution of morality; the subtitle of their book is *The Evolution and Psychology of Unselfish Behavior*. The first part of their book establishes the viability of altruism in the evolutionary sense ("fitness-sacrificing behavior" might be a better term), and the second part more tentatively argues that for cognitively sophisticated creatures like us, it is plausible that altruism in the vernacular, psychological sense is a proximate mechanism that natural selection might have struck upon for getting us to act in an appropriate fitness-sacrificing way.. But, as I argued earlier, creatures who are altruistic (psychologically), though perhaps "moral" in the sense of deserving praise, are not necessarily moral in the sense of evaluating themselves and each other in moral terms. Psychological altruism may correctly be called a "moral sentiment," but this just draws attention to the fact that creatures with no cognitive ability to grasp a moral concept or make a moral judgment can be ascribed a moral sentiment. If we're interested in the origins of moral *judgment*, then Sober and Wilson do not offer a theory. This is not a criticism of them, just an observation of what they do and what they do not attempt. Indeed, they are perfectly explicit about this, denying two theses: "that morality always requires us to sacrifice self-interest for the sake of others … [and] that to be motivated by an altruistic desire is the same thing as being motivated by a moral principle" (1999, p. 237).

2) However, though Sober and Wilson do not attempt it, it is perfectly possible that biological group selection could produce the trait of making moral judgments. If moral judgment reinforces prosocial behavior, then (*ceteris paribus*) it will be good for a group to contain members able and disposed to engage in moral thinking. However, it should be noted that general references to "prosociality" are rather coarse grained, and there is probably a more detailed story to be told about the characteristic subject matter of morality. A number of comprehensive cross-cultural studies have unanimously found certain broad universals in moral systems: (1) negative appraisals of certain acts of harming others, (2) values pertaining to reciprocity and fairness, (3) requirements concerning behaving in a manner befitting one's status vis-à-vis a social hierarchy, and (4) regulations clustering around bodily matters (such as menstruation, food, bathing, sex, and the handling of corpses) generally dominated by concepts of purity and pollution (see Haidt and Joseph, 2004, for discussion and references). The first three qualities all pertain directly or indirectly to reciprocal exchanges. (To see how indirect reciprocity might produce an emphasis on social hierarchy, recall the importance of reputation to such exchanges.) Given this, we may conclude that if the human moral sense is prepared for any particular subject matter it is surely

reciprocity; it therefore seems eminently reasonable to assume that reciprocal exchanges were a central evolutionary problem that morality was designed to solve. Saying this doesn't knock the other processes out of the running. Group selection—most probably at the cultural level—may well have also been a major factor. But my hunch is that reciprocity, broadly construed, is what got the ball rolling. (The moralization of disgust—giving rise to taboos concerning food and sex, for example—I suspect of being a matter of natural selection co-opting a motivational mechanism that had conveniently evolved for other initial purposes.)

There is also a body of evidence, alluded to earlier, suggesting that many of the concomitant traits one might expect would evolve in order to govern reciprocal exchanges are indeed innate features of human psychology: the interest in acquiring knowledge of others' reputations and in advertising one's own good reputation, our sensitivity to issues of distributive fairness in exchanges, our capacity to distinguish between accidental and purposeful harms (and our inclination to forgive the injuries of the former kind), our sensitivity to cheats and our antipathy towards them (our eagerness to punish them even at material cost to ourselves), and our heightened sense of possession. The crucial question is whether a moral sense forged by group selection could be expected to exhibit the same attributes. And I confess to finding this a very difficult question to assess. It is not obvious, for example, that group interests are served by members having elevated the possession relation into the moralized notion of *ownership*. It is not obvious that group interests will be served by members being acutely aware of distributive fairness—after all, *the group* might do just fine, or better, with a terribly inequitable and undeserved distribution of resources. Of course, saying that it is not obvious doesn't mean it's false. But it is reasonable, I think, at least to conclude that certain features that seem very central to morality fall smoothly and easily out of the "reciprocity hypothesis," but follow only with work from the group selection hypothesis. Hardly a decisive consideration, but a worthwhile dialectical point nonetheless.

What if it turns out that the two hypotheses equally well explain the available evidence? Then, by Sober and Wilson's own methodological lights, we should plump for the explanation in terms of individual selection (1999, p. 126). With careful reservations, they endorse George Williams's principle of parsimony that "one should postulate adaptation at no higher a level than is necessitated by the facts" (1966, p. 262). Their corollary is that "this does not allow one to reject a multilevel selection hypothesis without consulting the data … Multilevel selection hypotheses must be evaluated empirically on a case-by-case basis, not *a priori* on the basis of a spurious global principle" (1999, p. 126). Quite so. By merely putting forward a hypothesis I don't take myself to have established anything in advance of empirical evidence, but it is good to have options on the table before we start digging.

3) Finally, I want to acknowledge, but reject as uninteresting, the possibility argued for by Sober and Wilson that reciprocal altruism is really just a special form of group selection, involving a group of two (in the case of a straightforward direct reciprocal relation). For Sober and Wilson the relevant notion of a group constituting a vehicle of selection is a *trait group*: a population of $n$ individuals (where $n > 1$) "that influence each other's fitness with respect to a certain trait but not the fitness of those outside the group" (1999, p. 92). Kim Sterelny (1996) has argued plausibly that there is a difference *in kind* between groups that satisfy the above criterion (including partners in reciprocal exchanges) and the "superorganisms" often used as

paradigmatic examples of group selection (including especially colonies of social insects). Examples of the latter category exhibit an extreme degree of cohesion and integration, their members share a common fate, and such groups possess adaptations that cannot be equivalently redescribed at the individual level (e.g., the tendency of newly hatched queens to kill their sisters). Such groups have as respectable a claim to being robustly objective vehicles of selection as do organisms. Concerning examples of the former category, by contrast, the decision to describe selection as occurring at the level of the group is a purely optional one, for this group-level description is equivalent to an individual-level description. Regarding this category, Sterelny (following Dugatkin and Reeve, 1994) advocates a pluralistic approach, where the only difference between preferring individuals or trait groups as the vehicle of selection—that is, of regarding the process as one of individual selection or group selection—is a heuristic one, depending "on our explanatory and predictive interests" (1996, p. 572).

Going along with Sterelny, I am willing to concede that, on a certain liberal understanding of what it takes to be a group, reciprocal relations may count as group-selected, or they can be equivalently described in terms of individual selection. Any debate on the matter, says John Maynard Smith, is not "about what the world is like … [but] is largely semantic, and could not be settled by observation" (1998, p. 639). But it is clear that there is a kind of group selective process which they are *not* an example of: what Sterelny calls "superorganism selection" (1996, p. 577). One could argue that human cooperative faculties (e.g., morality) are the product of superorganism selection, or one might instead argue that they may be explained by invoking only, say, reciprocity. These are quite distinct hypotheses, and it cannot be reasonably denied that if we were unable to distinguish between them, due to a methodological decision to lump reciprocity (along with kin selection and the extended phenotype) under the umbrella term of "group selection," this would be an unacceptable loss of explanatory detail in the service of theoretic unification.

**References:**

Aiello, L. and Dunbar, R. (1993). Neocortex size, group size, and the evolution of language. *Current Anthropology, 34*.

Ainslie, G. (1992). *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. Cambridge University Press.

Alexander, R. (1987). *The Biology of Moral Systems*. Aldine de Gruyter.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*.

Bandura, A., Barbaranelli, C., Caprara, G.V. and Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*.

Bateson, P. (1991). Are there principles of behavioural development? In P. Bateson (ed.), *The Development and Integration of Behaviour*. Cambridge University Press.

Beer, J.S., Heerey, E.A., Keltner, D., Scabini, D. and Knight R.T. (2003). The regulatory function of self-conscious emotion: Insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology, 85*.

Boyd R. and Richerson, P.J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13.*

Bronson, W. (1975). Developments in behavior with age-mates during the second year of life. In M. Lewis and L. Rosenblum (eds.), *Friendship and Peer Relations*. Wiley.

Brown, D.E. (1991). *Human Universals*. Temple University Press.

Brown, R. (1973). *A First Language: The Early Stages*. Harvard University Press.

Covert, M.V., Tangney, J.P., Maddux, J.E. and Heleno, N.M. (2003). Shame-proneness, guilt-proneness, and interpersonal problem solving: A social cognitive analysis. *Journal of Social and Clinical Psychology, 22*.

Darwin, C. [1879] 2004. *The Descent of Man, and Selection in Relation to Sex*. Penguin Books.

Darwin, F. (ed.) (1887). *The Life and Letters of Charles Darwin*. Vol. 2. John Murray.

Dawe, H. (1934). An analysis of two hundred quarrels of preschool children. *Child Development, 4*.

Dennett, D.C. (1995). *Darwin's Dangerous Idea*. Simon and Schuster.

Dugatkin, L. and Reeve, H. (1994). Behavioral ecology and levels of selection: dissolving the group selection controversy. *Advances in the Study of Behavior, 23*.

Dugatkin, L. (1999). *Cheating Monkeys and Citizen Bees*. Harvard University Press.

Dunbar, R. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16*.

Dunbar, R. (1996). *Grooming, Gossip, and the Evolution of Language*. Harvard University Press.

Elster, J. (1984). *Ulysses and the Sirens*. Cambridge University Press.

Ferguson, T.J., Stegge, H., Miller, E.R. and Olsen, M.E. (1999). Guilt, shame, and symptoms in children. *Developmental Psychology, 35*.

Fisher, R. [1930] (1999). *The Genetical Theory of Natural Selection*. Oxford University Press.

Fiske, A.P. (1991). *Structures of Social Life*. Free Press.

Foot, P. (1958). Moral arguments. *Mind, 67*.

Griffiths, P. (2002). What is innateness? *Monist, 85*.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108.*

Haidt, J. and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Dædalus, 133*.

Heinsohn, R. and Packer, C. (1995). Who will lead and who will follow? Complex cooperative strategies in group-territorial lions. *Science, 269*.

Hollos, M., Leis, P.E. and Turiel, E. (1986). Social reasoning in Ijo children and adolescents in Nigerian communities. *Journal of Cross Cultural Psychology, 17*.

Kant, I. [1783] (2002). *Groundwork for the Metaphysic of Morals*. A. Zweig (trans.). Oxford University Press.

Keltner, D. (2003). Expression and the course of life: Studies of emotion, personality, and psychopathology from a social-functional perspective. In P. Ekman, J.J. Campos, R.J. Davidson and F.B.M. de Waal (eds.), *Emotions Inside Out: 130 Years After Darwin's "The Expression of the Emotions in Man and Animals." Annals of the New York Academy of Sciences, 1000*.

Keltner, D., Moffitt, T.E. and Stouthamer-Loeber, M. (1995). Facial expressions of emotion and psychopathology in adolescent boys. *Journal of Abnormal Psychology, 104*.

Ketelaar, T. and Au, W.T. (2003). The effects of guilty feelings on the behavior of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion, 17*.

Lahti, D.C. (2003). Parting with illusions in evolutionary ethics. *Biology and Philosophy, 18.*

Markessini, J. and Golinkoff, R. (1980). 'Mommy sock': The child's understanding of possession as expressed in two-noun phrases. *Journal of Child Language, 7*.

Maynard Smith, J. (1998). The origin of altruism. *Nature, 393*.

McBrearty, S. and Brooks, A.S. (2000). The revolution that wasn't: A new interpretation of the origin of modern human behavior. *Journal of Human Evolution,* 39.

Mellars, P. (1995). *The Neanderthal Legacy: An Archaeological Perspective from Western Europe*. Princeton University Press.

Nowak, M. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393.*

Nucci, L.P., Turiel, E. and Encarnacion-Gawrych, G.E. (1983). Social interactions and social concepts: Analysis of morality and convention in the Virgin Islands. *Journal of Cross Cultural Psychology, 14*.

Richerson, P., Boyd, R. and Henrich, J. (2003). Cultural evolution of human cooperation. In P. Hammerstein (Ed.), *The Genetic and Cultural Evolution of Cooperation*. MIT Press.

Roberts, S. (1979). *Order and Dispute: An Introduction to Legal Anthropology*. St. Martin's Press.

Rozin, P., Haidt, J., Imada, S. and Lowery, L. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76.*

Schelling, T.C. (1980). The intimate contest for self-command. *The Public Interest, 60.*

Smetana, J.G. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52*.

Smetana, J.G. and Braeges, J.L. (1990). The development of toddlers' moral and conventional judgments. *Merrill-Palmer Quarterly, 36*.

Smith, E. (2003). Human cooperation: perspectives from behavioral ecology. In P. Hammerstein (ed.), *The Genetic and Cultural Evolution of Cooperation.* MIT Press.

Smith, P. and Green, M. (1975). Aggressive behavior in English nurseries and play groups: sex differences and response of adults. *Child Development, 46.*

Sober, E. (1988). What is evolutionary altruism? In M. Matthen and B. Linsky (eds.), *Philosophy and Biology: Canadian Journal of Philosophy, suppl. vol. 14*.

Sober, E. and Wilson, D.S. (1999). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Song, M.J., Smetana, J.G. and Kim, S.Y. (1987). Korean children's conceptions of moral and conventional transgressions. *Developmental Psychology, 23*.

Spiro, M. (1958). *Children of the Kibbutz*. Harvard University Press.

Sripada, C.S. (2005). Punishment and the strategic structure of moral systems. *Biology and Philosophy, 20.*

Sterelny, K. (1996). The return of the group. *Philosophy of Science, 63.*

Tangney, J.P. (2001). Constructive and destructive aspects of shame and guilt. In A.C. Bohart and D.J. Stipek (eds.), *Constructive and Destructive Behavior: Implications for Family, School, and Society*. American Psychological Association.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46.*

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Turiel, E. (1998). The development of morality. In W. Damon (ed.), *Handbook of Child Psychology*, vol. 3. 5th edition. John Wiley & Sons, Inc..

Turiel, E., Killen, M. and Helwig, C.C. (1987). Morality: Its structure, functions, and vagaries. In J. Kagan & S. Lamb (eds.), *The Emergence of Morality in Young Children*. University of Chicago Press.

de Waal, F.B.M. and Luttrell, L. (1988). Mechanisms of social reciprocity in three primate species: symmetrical relationship characteristics or cognition. *Ethology and Sociobiology, 9.*

Williams, G. (1966). *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton University Press.

Yau, J. and Smetana, J.G. (2003). Conceptions of moral, social-conventional, and personal events among Chinese preschoolers in Hong Kong. *Child Development, 74*.

Zahavi, A. and Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford University Press.