Moral Reputation: An Evolutionary and Cognitive Perspective

DAN SPERBER AND NICOLAS BAUMARD

Abstract: From an evolutionary point of view, the function of moral behaviour may be to secure a good reputation as a co-operator. The best way to do so may be to obey genuine moral motivations. Still, one's moral reputation maybe something too important to be entrusted just to one's moral sense. A robust concern for one's reputation is likely to have evolved too. Here we explore some of the complex relationships between morality and reputation both from an evolutionary and a cognitive point of view.

People may behave morally because they intrinsically value doing so—a genuine moral reason—or in order to gain the approval of others—an instrumental reason. Both moral and reputational concerns are commonly involved in moral behaviour and cannot be pried apart without understanding their intricate relationships. Here we aim at contributing to such an understanding by investigating the role, content, and mechanisms of moral reputation.

1. Function and Motivation of Moral Behaviour

Throughout their lifetime, humans depend for their survival and welfare on frequent and varied cooperation with others. In the short run, it would often be advantageous to cheat, that is, to take the benefits of cooperation without paying the costs. Cheating however may seriously compromise one's reputation and one's chances of being able to benefit from future cooperation. In the long run, co-operators who can be relied upon to act in a mutually beneficial manner are likely to do better in what may be called the 'cooperation market.' According to a standard evolutionary approach to morality that we may call the 'mutualistic approach', the biological function of moral behaviour is precisely to help individuals gain a good reputation as co-operators (Alexander, 1987; Krebs, 1998; Trivers, 1971).[1]

We are grateful to Deirdre Wilson and to four anonymous reviewers for their very useful comments and suggestions.

Address for correspondence: Dan Sperber Institut Jean Nicod, ENS 29 rue d'Ulm, 75005 Paris, France.
Email: dan.sperber@gmail.com

[1] This 'mutualistic' approach has been challenged by scholars who view morality as having primarily evolved through group rather than individual selection and for the benefit of the group rather than of individuals (e.g. Sober and Wilson, 1998; Gintis *et al.*, 2005). In this second 'altruistic' approach to morality, reputation plays a less central role than in the mutualistic approach, on which we focus. We do however briefly compare the two approaches in Section 2. For a comprehensive discussion of the evolution of morality and a defence of the mutualistic approach, see Baumard, 2010; Baumard, André and Sperber, forthcoming.

Mind & Language, Vol. 27, No. 5 November 2012, pp. 495–518.

If the function of moral behaviour is to safeguard one's reputation, then shouldn't one act morally when one's reputation is at stake, and selfishly otherwise? As Hume puts it (attributing the thought to a 'sensible knave'): 'honesty is the best policy, may be a good general rule, but is liable to many exceptions; and he [. . .] conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions' (Hume, 1898, p. 257). From a mutualistic point of view then, morality is problematic. It is no doubt advantageous to have others behave morally but, or so it seems, it is no less advantageous to merely *appear* to behave morally oneself. Isn't then an evolved mutualistic morality essentially Machiavellian? Not necessarily.

The biological function of a given type of behaviour need not coincide with its psychological motivation. After all, the function of sexual intercourse is procreation, but relatively few instances of sexual intercourse (and none among earlier humans and animals unaware of the biology of reproduction) are motivated by the desire to procreate. So, even if the *function* of moral behaviour is to secure a good reputation, this leaves open the possibility that its *motivation* is more genuinely moral. In his landmark paper on 'reciprocal altruism'—the first modern mutualistic sketch of the evolution of morality—Trivers (1971) drew on this classical distinction between biological function and mechanism, and suggested as an aside that, in a number of cases, the function of securing a good reputation would be better served by behaviour based on genuinely moral dispositions than by the direct pursuit of reputational gains:

> Selection may favour distrusting those who perform altruistic acts without the emotional basis of generosity or guilt because the altruistic tendencies of such individuals may be less reliable in the future. One can imagine, for example, compensating for a misdeed without any emotional basis but with a calculating, self-serving motive. Such an individual should be distrusted because the calculating spirit that leads this subtle cheater now to compensate may in the future lead him to cheat when circumstances seem more advantageous (because of unlikelihood of detection, for example, or because the cheated individual is unlikely to survive) (Trivers, 1971, p. 51).

A comparable account of the moral emotions has been developed by Robert H. Frank (Frank, 1988, 2004; see also Hirshleifer, 1987). Frank proposes a two-step account of the evolution of these emotions. A central idea is that emotions, being hard to fake, provide an honest signal; moral emotions in particular provide an honest signal of trustworthiness. As argued by Niko Tinbergen (1952) however, biological signals cannot originate as such. They first have to evolve with a different function or as by-products before acquiring the function of signalling. In the case of the moral emotions, Frank suggests that they first evolved to guide our behaviour and control our impulse to achieve immediate reward. In a recent formulation:

> If you were endowed with a moral sentiment that made you feel bad when you cheated your partner, even if no one could see that you had that sentiment,

> this would make you better able to resist the temptation to cheat in the first round. And that, in turn, would enable you to generate a reputation for being a cooperative person, clearly to your advantage.
>
> Moral emotions may thus have two separate roles. They are impulse control devices. The activation of these emotions, like other forms of brain activation, may be accompanied by involuntary external symptoms that are observable. If so, the observable symptoms over time become associated in others' minds with the presence of the moral sentiments. And once that association is recognized, the moral emotions are able to play a second role—namely, that of helping people solve one-shot prisoner's dilemmas [i.e. dilemmas presented by a situation where it would be beneficial if all the people involved in a one-shot interaction cooperated, but where it is even more advantageous for each individual to cheat]. Then symptoms themselves can then be further refined by natural selection because of their capacity to help other identify people who might be good partners in one shot dilemmas (Frank, 2006, p. 202).

Both Trivers and Frank focus almost exclusively on the role of moral emotions and give little attention to elaborating and defending what, we believe, has to be the cognitive core of any such approach. We are not denying that moral emotions play a major role in morality and that a full account of the evolution of morality must give them a central place. Still, moral emotions are moral because of their cognitive content. To feel guilty or indignant involves judging an action to be morally wrong—an action of one's own in the case of guilt, of someone else in the case of indignation. Moral emotions could not evolve without a capacity for moral judgment evolving too. Moreover, for moral emotions to signal trustworthiness, as argued by Trivers and Frank, they must be based on effective recognition and appropriate evaluation of actions that make one a trustworthy or on the contrary an untrustworthy partner in cooperation. In such an evolutionary perspective, moral content cannot vary arbitrarily: it has to correspond to the kind of behaviour and qualities that foster mutually rewarding cooperation.

Imagine two individuals, one with overwhelming moral emotions based on poor, generally self-serving moral judgment, and another with low-key moral emotions based on good and impartial moral judgment. Which one would you prefer as a partner? Moral emotions are a signal of trustworthiness only to the extent that the underlying moral judgment is itself trustworthy and actually guides the individual's behaviour. In enduring social relationships, the best evidence of an individual's moral judgment and dispositions is given not by her emotions but by her behaviour over time. In the ancestral environment where moral dispositions evolved, most, if not all, social relationships were enduring ones. In a modern environment where having to interact with strangers can be a daily experience and where the 'one-shot dilemmas' Frank talks about are relatively common, the expression of moral emotions may indeed play an important role in moral assessment. Even in this modern environment however, a person's moral emotions are first and foremost a signal of her attitude to an on-going situation or relationship and the perception of

these emotions is relevant to deciding how to interact with her in this situation or relationship. This coordinating role, we would suggest (following Gibbard 1990), might well be the main function of moral emotions as signals: they are calls for a direct and specific response whatever light they may also throw on the moral trustworthiness of the person.

Emotions are honest signals, Frank argues, because they are hard to fake. This is indeed plausible and relevant to the role of emotions both in signalling moral character and in signalling moral attitude to a situation or a relationship. This however raises an interesting issue. In McElreath's words:

> [W]hy would natural selection not favor individuals who could fake emotional displays and therefore exploit cooperators? One possibility is that the production of emotional displays is physiologically costly. However, no careful and accepted argument exists as to why this might be the case (McElreath, 2003, p. 145).

A possible answer to McElreath's challenge is provided by Frank's own suggestion that the initial function of moral emotions may have been to guide our behaviour towards longer-term benefits when our impulse might be to seek immediate gratification. Respect for others' right stands in the way of greed, indignation stands in the way of selfish indifference. If so, feigning emotions might occasionally increase the benefits of signalling but at the cost of forsaking the benefits of behaviour guidance, with the risk that one's deeds would soon belie the expression of one's emotions. As many parents know, faking the facial and bodily expression of anger without actually losing your temper fails to frighten children into submission not because they immediately see that you are dissembling, but because your tepid behaviour soon alleviates the fear instilled by your apparent heat.

Moral emotions can themselves be quite impulsive and at times quite costly, as when indignation causes one to overreact. Hence, while it is quite plausible that the primary function of moral emotions is to guide behaviour, it is less evident that they would do so by being particularly apt at controlling impulse. Wouldn't a Machiavellian disposition involve not only the ability to master one's impulses at least as much as would moral emotions, but also a better ability to anticipate costs and benefits? Just as some moral emotions come with a sense of urgency, some selfish aims are best pursued with self-discipline: ask a swindler! For gaining a reputation as a good co-operator, being guided by moral emotions is clearly better than being guided by the search for immediate gratification, but mightn't being guided by a Machiavellian disposition be better still?

Actually, a Machiavellian strategy of behaving morally when one's reputation is at stake and selfishly otherwise carries cognitive costs and practical risks. From a practical point of view, an error—for instance, mistakenly assuming that no one is paying attention to a blatantly selfish action—may compromise an agent's reputation (see Delton, Krasnow, Cosmides and Tooby, 2011 for instance). Such a mistake may not only cause direct witnesses to lower their opinion of the agent,

but is also likely, given the typically human way of spreading information, to influence many more people.

From a cognitive point of view, a Machiavellian strategy is a demanding one. It is often difficult to tell whether others are paying attention to our behaviour, and to predict how they might interpret it and what they would think or say about us as a result. Even if a Machiavellian agent cleverly manages to avoid being caught cheating, she might still behave in a way that suggests she is being clever rather than moral, and compromise her reputation as a result.

A number of studies in behavioural economics confirm that it is not that easy to pretend to be genuinely moral. Participants in experiments are able to predict in advance whether or not their partners intend to cooperate (Brosig, 2002; Frank, Gilovich and Regan, 1993). They base their judgments on the likely motivations of others (Brehm and Cole, 1966; Schopler and Thompson, 1968), on the costs of their moral actions (Ohtsubo and Watanabe, 2009), or on the spontaneity of their behaviour (Verplaetse, Vanneste and Braeckman, 2007). More generally, many studies suggest that it is difficult to completely control the image one projects, and that there are numerous indirect cues beside manifest emotions to an individual's propensity to cooperate (Ambady and Rosenthal, 1992; Brown, 2003).

Further studies confirm that people are quite good at evaluating the moral dispositions of others. Pradel and colleagues (Pradel *et al*., 2009), for example, had middle school and high school classes play a dictator game (where an individual may share an endowment or keep it for herself). Students were then asked to guess how their classmates had behaved. The pupils were biased in favour of their friends, but they were still quite good at guessing how generous their classmates had been. Moreover, pupils who behaved more morally in the game tended to choose each other as friends. These results suggest that cooperativeness can be detected, and that people who are more cooperative tend to choose one another as partners (see also Sheldon *et al*., 2000).

Machiavellian strategies for securing a good moral reputation without paying the cost of morality are thus both hard to follow and risky. Is there a way cognitively easier and safer than a Machiavellian strategy to secure such a reputation? Yes: it consists in deserving it, that is, in having a genuine, non-instrumental preference for moral behaviour and a disposition to act on the basis of this preference. At the cost of missing a few opportunities for profitable cheating, a genuinely moral person is in a uniquely good position to be regarded as such.

Trivers' suggestions and Frank's more elaborate ideas on the role of moral emotions deserve more discussion and empirical investigation than have received, but, by raising many further issues, they blur the core idea that the best way to acquire a moral reputation may be to behave morally. Doing so provides the strongest and most reliable evidence of one's moral character, whatever the exact role emotions might play in highlighting and reinforcing this evidence.

In fact, relatively little attention has been paid to Trivers' suggestion that while the function of moral behaviour is to gain a good moral reputation, this function may be best served by a genuine moral motivation (but see Delton *et al*., 2011).

As we explain in the next section, this suggestion (made more vulnerable to criticisms because of being entwined with speculations on the role of emotions) has been overshadowed by the debate on Trivers' central notion of reciprocal altruism.

2. Reputational Concerns

It is now generally agreed that 'reciprocal altruism' as initially proposed, even when taken together with Hamilton's (1964) explanation of kin altruism, falls short of providing an adequate explanation of human cooperation and morality. In non-human animals, the conditions for the evolution of reciprocal altruism are hardly ever met; the extent to which they are met in the human case is contentious. Moreover people commonly help others even when they have no reason to expect direct reciprocation.

Trivers developed the idea of reciprocal altruism to help explain how cooperation can evolve among non-kin through individual-level selection. He did so in the 1960s at a time when the idea of group-level selection, as expounded in particular by Wynne-Edwards (Wynne-Edwards and Copner, 1962) and criticised by Maynard Smith (1964) and Williams (1966), seemed quite discredited. Since then, revised and more plausible notions of group selection have however been proposed by D. S. Wilson (1975) and others. They have led to the development of influential alternative approaches to morality that give a main role to group selection in the evolution of altruism (genetic evolution, as in Sober and Wilson, 1998; or both genetic and cultural evolution as in Gintis *et al*. 2005). The kind of altruistic behaviour that would evolve under group selection has as its *function* to benefit others rather than the individual actor. Securing the individual's reputation is not indifferent, but it is not central. One need not appeal, in such an altruistic perspective, to the function/mechanism distinction in order argue that humans have a truly moral disposition.[2]

Others scholars, sceptical of the role of group selection, have tried to improve the mutualistic approach by drawing on the resources of evolutionary psychology, by generalising the idea of reciprocal altruism to so-called indirect reciprocity (Alexander, 1987; Milinski, Semmann and Krambeck, 2002; Nowak and Sigmund, 2005). Like Trivers, these scholars see the function of moral behaviour as that of providing individuals with a reputation as reliable co-operators. Unlike Trivers however, most of these scholars explicitly assume that, at the psychological level, moral behaviour is typically guided by reputational concerns, suggesting that the main if not the only motivation for behaving morally is self-interested and

[2] Of course, from a normative point of view, being disposed to favour one's own group even at the expense of other groups may not quite qualify as a 'moral' disposition, but this is beside the present point.

instrumental rather than genuinely moral. Bateson, Nettle and Roberts (2006), for instance, write:

> Our results therefore support the hypothesis that reputational concerns may be extremely powerful in motivating cooperative behavior. If this interpretation is correct, then the self-interested motive of reputation maintenance may be sufficient to explain cooperation in the absence of direct return (Bateson, Nettle and Roberts, 2006, p. 413).[3]

Much of the debate on the evolution of morality has revolved around the evidence provided by the use of experimental paradigms developed by behavioural economists. These 'economic games' have become a central tool in the study of human cooperation and morality. The evidence they provide seems at first sight to imply that people are genuinely moral after all. In the 'dictator game' for instance, one participant (the 'dictator') receives a certain amount of money and must decide how much of it to give to a second participant. The game generally takes place in conditions of strict anonymity. The dictator faces a computer screen, does not meet the second participant (who plays a purely passive role), and is told (truthfully) that not even the experimenter will know how much she will have chosen to give. Hundreds of experiments across the world have shown that first participants, be they American students or Amazonians foragers, do not behave in a strictly selfish fashion. On average, more than 60% of dictators give money, with the mean transfer being roughly 20% of the endowment. In a debate where the two alternatives seemed to be either that there exists genuine altruism and it has been selected at the group level, or that 'moral' dispositions have been selected at the individual level and are actually self-interested, this has been taken as strong evidence in favour of the group-selection account of altruism.

Is this evidence from behavioural economics enough to falsify the idea that, even when they seem to behave altruistically, people are actually pursuing their own selfish interest? Not really. It may be almost impossible to completely eliminate the apprehension that others might somehow be informed of one's action, so that, even under these conditions of anonymity, participants might still be influenced by concern for their reputation. It is conceivable that participants suspect, even if vaguely and unconsciously, that they stand to gain by being more generous than required by the rules of the game, or that they stand to lose by being strictly selfish. Might this explain why dictators give some money,

[3] Similar claims are made by Andreoni and Petrie, 2004; Barclay, 2004; Bereczkei, Birkas and Kerekes, 2007; Bourrat, Baumard and McKay, 2011; Burnham, 2003; Burnham and Hare, 2007; Ernest-Jones, Nettle and Bateson, 2011; Haley and Fessler, 2005; Hardy and Van Vugt, 2006; Hoffman, McCabe and Smith, 1996; Kurzban, 2001; Kurzban, DeScioli and O'Brien, 2006; Mifune, Hashimoto and Yamagishi, 2010; Milinski *et al*., 2002; Piazza and Bering, 2008a, 2008b; Rigdon, Ishii, Watabe and Kitayama, 2009; Van Vugt and Hardy, 2010).

even though they are under no obligation to do so and would be better off keeping it all to themselves? This is what has been argued by psychologists and evolutionary theorists who defend mutualistic accounts of the evolution of human cooperation.

A good example of this 'reputational concern' approach is provided by a set of clever experiments conducted by Haley and Fessler (2005), who manipulated factors that might increase or decrease participants' concern for their reputation. In one condition, for instance, the desktop background displayed two stylized eye-like shapes and more participants gave money than in the control condition (where, instead of the eyes, the desktop background displayed the lab's logo). In another condition, in contrast, participants were fitted with noise-cancelling earmuffs: in this case, fewer of them gave. These findings suggest that eyes and certain noises are unconsciously processed by the brain as cues to a social situation where reputation is at stake, causing more participants to be generous.

The findings of Haley and Fessler—as well as similar findings by, among others, Bereczkei, Birkas and Kerekes (2007); Hoffman, McCabe and Smith (1996); and Kurzban (2001)—raise the issue of the degree to which apparently moral behaviour is in fact guided by self-interest. They show that people are more likely to forsake short-term gains when there are cues suggesting that they might be observed and that their behaviour may be found objectionable. This seems to provide evidence against the view that participants in economic games give money out of a genuine altruistic disposition resulting from group selection and in favour of the view that seemingly altruistic actions are in fact governed by a reputational concern.

Trivers' suggestion that while the function of morality is reputational, its motivation may be genuinely moral is probably unattractive to both sides of this debate. It is unattractive to group-selection theorists because they see the function and not just the motivation of morality as being genuinely altruistic. It is unattractive to individual-selection theorists, because the hypothesis that people have a genuine moral motivation does not entail the prediction that participants should be sensitive to reputational considerations; hence, if true, it would detract from the evidential value of results such as Haley and Fessler's in this debate.

The evidence we have just considered shows that people are influenced by reputational concern in economic games. This evidence may be seen as casting doubt on the genuinely moral character of people's motivation, and hence as weighing against the group-selection altruistic approach to morality and in favour of the individual-selection mutualistic approach stripped of Trivers' suggestion that the best way to acquire a good moral reputation may be to be genuinely moral in one's motivations. Still, this is hardly crucial evidence sufficient to settle the issue.

What would count as evidence *against* the view that moral behaviour is guided by a concern for reputation? Suppose the real goal of apparently moral actions were to safeguard one's reputation. In this case, one should never be tempted to act morally when the obvious effect of such an action would be to damage it.

We know of no experiment exploring such a possibility.[4] On the other hand, we all are acquainted with, or have heard of, people who have acted morally at a cost to their reputation. This happens, for instance, when a person confesses out of a sense of guilt to some serious moral violation that nobody had suspected. It also happens when one refuses to take part in a misdeed at the cost of losing the esteem of one's friends, not for fear of being caught but from sheer moral disapproval. A striking example of an individual acting in such a manner at a very serious cost to his reputation is that of the Nobel Prize-winning German writer Günter Grass. Grass, who was a left-wing moral authority and often had faulted his country for not fully facing its Nazi past, admitted in 2006 that, for more than forty years, he had hidden the fact that he himself had been, at the age of seventeen, a member of the Waffen-SS. Grass' behaviour in revealing his Nazi past would be quite hard to interpret without assuming that it was inspired by genuine moral concerns.

Haley and Fessler's evidence suggests that apparently moral behaviour is governed by a concern for one's reputation. Common anecdotal evidence and cases such as that of Günter Grass suggest that people may on occasion override this concern and act on genuinely moral grounds. How can we reconcile and articulate what is true and relevant in both kinds of evidence?

3. Different Motives, Same Behaviour

Assume that people have a genuinely moral disposition. Even so, this disposition would not be the sole basis of their choices and behaviour. People have aesthetic and prudential dispositions and are, needless to say, not above also being motivated by the pursuit of self-interest. A single set of available alternatives can activate several systems of preferences, resulting in either a compromise or in one or other of these systems carrying the decision. For example, dietary choices can be influenced at once by taste and by health preferences. These two types of preferences can be mutually reinforcing (as in the case of fruits and vegetables) or in conflict (as with 'junk food'). That people have a 'plurality of passions' is not an original remark. Butler, for instance, noted that anger and affection for one's children, two natural passions, may come into conflict, and that a person 'may follow one passion and contradict another' (Butler, 1726/2008).

Just as it is plausible that there has been selective pressure for an authentically moral disposition whose function is to contribute to the individual's moral reputation, it

[4] Designing behavioural economics experiments to test this hypothesis is particularly challenging, if not impossible. This would involve causing a first participant to believe that, if she behaved morally, a second participant would misinterpret her behaviour and have a negative opinion of her as a result. This cannot be done without misinforming at least one participant. Misinforming participants however is not allowed in behavioural economics.

is plausible that there has been selective pressure for a disposition to be directly concerned about the effects of one's actions on one's reputation (with the two dispositions nonetheless being distinct). Concern for reputation has long been of interest to philosophers, psychologists and sociologists. They have described, on the one hand, the often complex strategies used to enhance reputation, and on the other, a spontaneous tendency to pay attention to the 'presentation of self in everyday life,' to use Goffman's famous title (1959). This concern is at work in every domain where our reputation can affect the success of our interactions with others: hence, with respect not only to moral qualities, but also to strength, intelligence, health, etc. When we may be being watched, we are motivated, consciously or not, to put ourselves in the best possible light. A comparable tendency to act so as to create a favourable image of oneself, for instance with respect to strength, can be found in many social animals (e.g. de Waal, 1989).

The distinction between a moral disposition and a disposition to attend directly to one's own reputation is particularly clear at the emotional level. Take shame and guilt. Shame typically occurs in situations where one's reputation might be jeopardised. Being ashamed causes one to flee the gaze of others and to try and downplay relevant facts. Guilt, in contrast, is a more specifically moral emotion. Feeling guilty causes one, when possible, to confess, to apologize, and to make amends (Tangney, Stuewig and Mashek, 2007). The more public an objectionable behaviour, the greater the associated shame, but not the associated guilt (Smith, Webster, Parrott and Eyre, 2002). Shame more than guilt goes with brain activity in areas involved in mindreading (Takahashi *et al*., 2004). Shame and guilt have different behavioural correlates. Here are a couple of experimental illustrations. In an experiment by Leith and Baumeister (1998), one group of participants had to recall an action that they felt guilty about, and a second group an action that they felt ashamed about. Participants in the 'guilt' condition were more likely to pay attention to the problems of others than those in the 'shame' condition, who were more likely to concentrate on their own problems. In a study by Ketelaar and Au (2003), participants who had to share money gave more when they had been asked to recall an action that they feel guilty about than when they had been asked to recount an ordinary day. No such effect was observed, on the other hand, when participants had been asked to remember a shameful event.

Of course, shame and guilt are mutually compatible and sometimes intimately associated. A single action can elicit both.

> Imagine for example someone who cheats or gives in to cowardice and then feels both guilty and ashamed. He feels guilty because he has acted contrary to his sense of right and justice. By wrongly advancing his interest, he has transgressed the rights of others, and his feelings of guilt will be more intense if he has ties of friendship and association to the injured parties. He expects others to be resentful and indignant at his conduct, and he fears their righteous anger and the possibility of reprisal. Yet he also feels ashamed because his

> conduct shows that he has failed to achieve the goal of self-command, and he has been found unworthy of his associates upon whom he depends to confirm his sense of his own worth. He is apprehensive lest they reject him and find him contemptible, an object of ridicule. In his behavior, he has betrayed a lack of the moral excellences he prizes and to which he aspires (Rawls, 1971, p. 445).

As H. B. Lewis (1971) writes, when we feel shame, we think '*I* did this horrible thing,' whereas when we are guilty, we think 'I did *this horrible* thing.' The action may be the same; its relevant features are not.

What we suggest then is that participants in the dictator game may be moved *both* by their moral sense *and* by their concern for their reputation. This provides an alternative interpretation of Haley and Fessler's results. 'Our results,' they write, 'challenge the claim that [. . .] prosocial behaviour in anonymous noniterated economic games cannot be explained with reference to reputational factors' (Haley and Fessler, 2005, p. 284). For them, cues to the presence of others are not completely eliminated in an experimental setting, whatever the assurance of anonymity given to the participants. It is these residual cues that cause participants to give, rather than an altruistic disposition resulting from group selection (as argued for instance by Fehr and Henrich, 2003). The alternative we propose is that cues to the presence of others activate a disposition to attend to one's reputation and add an extra non-moral motivation to act morally. This same reputational disposition may result, we hypothesise, not only in more 'moral' behaviour, but also in better control of one's posture and greater efficiency in one's movements. In other situations, it would result in greater speed, in uses of greater strength, in better linguistic performance and so on.

This hypothesis is experimentally testable. It predicts improvements in performance not only in the moral domain but also in all socially relevant domains, with two possible types of explanation. Either the dispositions governing performance in each of these domains are targeted at establishing and reinforcing reputation, or else an independent concern for reputation influences performance in all domains where reputation may seem to be at stake. If the second explanation is correct, then, by using cues as cognitively low-level as those used by Haley and Fessler (pictures of eyes or noise elimination), one should be able to obtain comparable performance improvements in a variety of non-moral domains. In fact, the psychological literature contains many relevant examples. Numerous experiments have demonstrated the effect of the experimenter's or third parties' presence on participants' performance. The more participants feel observed, the better their results. This phenomenon is known as the 'mere presence effect' (for a review, see Guerin, 1986). The explanation of Haley and Fessler's results in terms of the interaction of two dispositions, one genuinely moral and the other reputational, undermines their relevance as evidence against the group-selection altruistic approach but heightens their relevance to the study of the interaction of these two dispositions in the guidance of behaviour.

4. First- and Third-Person Morality

How do morality and reputation interact in evolution? We use moral judgment not only to guide our own behaviour but also to evaluate the behaviour of others. This third-person evaluative function of moral judgment is usually taken to be secondary and to derive from its first-person behaviour-guiding function. What is evaluated, after all, is the degree to which this behaviour-guiding function has been fulfilled in a given individual's behaviour. From a mutualistic evolutionary point of view however, the evaluative functions or morality must be at least on par with its guiding functions; if anything, the evaluative function should be seen as primary. Here is why.

If morality has the function of helping individuals gain a good reputation as co-operators, then this cannot be its only function. Reputation, more than anything else, is 'in the eye of the beholder' or more exactly in this case, in the words of people who express an opinion on someone else (for a more precise account of what is involved in reputation, see below). In other terms, if the function of moral behaviour is to secure reputation, then it is an adaptation to others' moral evaluation. But for this to be the case, moral evaluation itself had to evolve. That much is implicit in reputational accounts of the function morality. Let's make it explicit.

Just as it is adaptive to act so as to be chosen as a partner in cooperative ventures, it is adaptive to be able to choose reliable individuals as partners. In fact, this evaluative function would be adaptive even if no one made any effort to acquire a good reputation, whereas if no one bothered to evaluate your performance, trying to acquire a good reputation would be useless. Imagine an earlier stage when, as is the case with other primates now (but see Brosnan and De Waal, 2003), there is cooperation but no morality, and individuals cooperate only to the extent that it is in their short-term interest to do so. Even so, individuals would differ in the regularity with which they cooperate. Some may cooperate more regularly just because they are not smart enough to see opportunities of advantageous defections. Others may see the opportunities but find some intrinsic reward in continuing a cooperative interaction once begun. Whatever the causes, differences there have to be. Now, when choosing among potential partners with whom to cooperate, taking these differences into account would be adaptive. But then, if individuals are choosy in selecting partners, this creates a selective pressure for behaving so as to be chosen.

What can be called the 'evolution of cooperation by partner choice' should be seen as a special case of 'social selection' (Baumard, André and Sperber, forthcoming; Dugatkin, 1995; Nesse, 2007; M. West-Eberhard, 1979; M. J. West-Eberhard, 1983; for a brief history of the idea of social selection, see Nesse, 2009). Social selection is a special case of natural selection that occurs when individuals are in competition for some scarce resources and where winning the competition depends on the social choices of other individuals. The best-known case of social selection is that of sexual selection. In typical sexual selection, males are competing for sexual access to females and females do the choosing. There is then a selective pressure to develop, to a greater degree than one's competitors, whatever trait appeals to females. This

may yield runaway processes that result in exaggerated and costly traits such as the peacock's tail; however high the costs of such a trait, it is outweighed, in strong males, by the benefit it provides in mate competition. The same general mechanism of social selection may apply to any form of competition, and in particular to the competition to be chosen as a partner in profitable cooperative ventures (be they hunting expeditions, business investments, or the writing of a joint article).

In choosing a partner, one looks for relevant competencies and resources (which differ from one type of venture to another) and for cooperativeness or fairness, that is, for a reliable disposition to share the costs and to refrain from taking more than one's share of the benefits of cooperation (a disposition which is relevant in nearly all cases). Predicting an individual's future behaviour just on the basis of her past behaviour would ignore psychological factors that, in the human case, are crucial. A mere behavioural assessment may be good enough in other animals' repetitive forms of mutualistic cooperation (as between cleaner fish Labroides dimidiatus with client reef fish—see Bshary and Schäffer, 2002). In the human case however, given the open-ended variety of forms and conditions of cooperation and the complexity of people's beliefs and motivations, cooperativeness cannot be effectively assessed without making inferences about others' mental states and dispositions.

The intuitive psychology that humans use in order to, among other things, assess potential partners' cooperativeness has two components: a first component that infers beliefs from desires and actions, infers desires from beliefs and actions, and predicts actions given beliefs and desires; and a second component that, given instances of beliefs-desires-actions patterns, infers psychological dispositions or 'character'. In choosing among potential partners, we infer how they are likely to behave in the future from the dispositions we attribute them.[5]

The social selection pressure on individuals competing in the cooperation market comes from other individuals who find themselves in a position where they can choose partners and where they do so with the kind of psychological mechanisms we have just sketched. To do well in such competition, people have therefore to provide through their actions evidence of their dispositions. As we argued in Section 2, to try to do so in a Machiavellian way—being cooperative when one's action may be observed, being selfish otherwise—is effort-demanding and risky. To do it 'naturally', that is, through an evolved moral disposition, is generally more effective.[6] Moreover, the more 'readable' a good moral disposition, the better.

[5] Character or disposition psychology is prone to what Ross (1977) has called the 'fundamental attribution error': attributing too much causal power to character and not enough to the situation. There are two reasons why character psychology may nevertheless be effective in partner choice. The first reason is that one is choosing among possible partners who would all be, if chosen, in the same situation. What would make the difference, then, are the competitors' dispositions. The other reason is that, in their role as competing potential partners, people develop easily identifiable dispositional traits that help their chances of being chosen by truly making them more predictable.

[6] This leaves room for the frequency-dependent selection of true embezzler skills in some people, who do well as long as they are a small minority mimicking the genuine honesty of most folks.

Hence a modicum of generosity over and above what fairness strictly requires —yielding rather than haggling when the exact fair shares are in doubt—may be a highly effective way of advertising one's fairness.

Nesse (2007) makes the interesting suggestion that social selection through partner choice would lead to runaway phenomena comparable to those found in sexual selection, with the development of generosity well beyond what fairness requires. We are not convinced, however. Suppose the cooperation market is fluid and there is competition both to be chosen as partner and, for people proposing a cooperative venture, to attract the best partners possible. Then the point of equilibrium may well be a fair distribution of costs and benefits among partners, whoever initiated the partnership: competitors won't look for proposers requesting more than a fair share and won't be chosen if they themselves offer less (for a formal argument to that effect see André and Baumard, 2011). This is why, whereas amazing tails are common among male peacocks, Mother Theresa types are quite rare among humans. Actually, however much you may admire a saintly person, she might not be your first choice as a partner in cooperation: her giving too much and asking too little would put your more balanced behaviour in a bad light and might cause you to feel embarrassed. Her duties being to humankind (or to god), she may at any time leave you flat in order to achieve a greater good. Better a partner who can be expected to act in mutually beneficial ways in part because of the benefit doing so brings her than one who might do more for you than you do for her but who, precisely because she is not serving her own interest, remains quite free to ignore yours. This subtle balance of expectations and commitments that is an intuitive consequence of a mutualistic disposition presents, on the other hand, a challenge to a Machiavellian, who, lacking the intuition, is likely to end up doing too little or too much. The cooperation market selects not for just any kind of morality, but for fairness.

Of course, in many social settings, in particular modern ones, the cooperation market is quite imperfect and lacks fluidity. As a result, less powerful individuals are forced to accept unfair conditions. What is remarkable however is that, in such conditions, unfairly treated people are typically aware of the unfairness (see Abu-Lughod, 1986; Neff, 2003; Turiel, 2002), suggesting that a fairness morality is an evolved universal human trait.

It is to a large extent the same people who find themselves sometimes in a position to choose partners, sometimes in the position of competing to be chosen, and sometimes in the position of simultaneously choosing and being chosen. The cognitive abilities and dispositions to do both have evolved in all humans and involve as a central ingredient, we have suggested, mechanisms that provide intuitions of fairness.[7]

[7] Note that, in evaluation of others, fairness is approached as a desirable virtue whereas in guiding one's own behaviour it is approached as deontic principles. From an evolutionary point of view, we all have simultaneously simple versions of what moral philosophers call a 'virtue ethics' and a 'deontology' that we deploy for different purposes.

However, we do more than choose partners and compete to be chosen. We communicate about others and contribute to building their moral reputation; we communicate about ourselves to protect our own. What cognitive abilities are involved in this moral communication?

5. Reflective Aspects of Morality and Reputation

So far, we have considered two mental mechanisms that may cause moral behaviour: an evolved mutualistic moral sense and an evolved concern for one's reputation. Both, we assumed, are intuitive mechanisms: that is, they influence behaviour without necessarily involving any thinking about the reasons for one's actions. The moral sense provides intuitions about what is right and wrong and involves moral sentiments with motivating power. The kind of concern for reputation we have discussed so far is a disposition to automatically improve one's performance when there are cues to the effect that one might be observed by others (and, arguably, any form of interaction intuitively provides such a cue).

Let us call thinking about one's own thoughts, and in particular about reasons for one's beliefs and decisions, 'reflective thinking' (Mercier and Sperber, 2009; Sperber, 1997). Obviously, reflective thinking may play a causal role in moral behaviour. In fact, what we have said so far suggests principled reasons for why reflection is relevant to morality and to the management of reputation.

We have argued that a mutualistic morality is as much about judging other's people behaviour as it is about guiding one's own. Of course, moral evaluation of others can be just as intuitive as moral self-guidance. We can experience moral indignation and moral admiration without articulating reasons to justify these sentiments. If the function of our moral behaviour were just to cause individual observers to form their own good opinion of our moral character, such intuitive evaluation on their part might suffice. But what is at stake is not just other people's opinion of us, but also our reputation, and the two are not equivalent. Often, particularly in economics and in evolutionary theory, 'reputation' is used as a quasi-synonym of 'opinion': to have a good reputation is to have the good opinion of others, whichever way this opinion was acquired. However, there are important theoretical reasons for distinguishing reputation from opinion. (The point is not terminological and those who might prefer to treat the two terms as synonyms would still have to make the relevant distinction in other terms.)

Reputation is not just any kind of opinion, not even any kind of shared opinion. Reputation is an important aspect of human sociality and culture. It is a socially transmitted typically evaluative judgment that is presented as consensual, or at least as widely shared. Reputation is typically spread through conversation, and in the modern world, through the mass media and now the Internet. The utterances that convey a reputation refer, explicitly or implicitly, to that reputation. Others may each have a good opinion of you, but this does not imply that you have a good reputation. Others may express a good opinion of you, but if each of them expresses

it as their own personal opinion, this still does not imply that you have a good reputation. What is needed is that this positive opinion be seen as one shared in a group or in a milieu, as 'common ground' (D. K. Lewis, 1969) or as 'mutually manifest' (Sperber and Wilson, 1995), that is, *as a reputation*. When conveying a person's reputation, one typically describes it as such, or one uses phrases such as 'she is said to be . . . ,' 'Apparently he . . . ', 'They are recognised as the best . . . ,' 'It got the best rating . . . ' and so on.

Non-human animals do not have conversations about each other, and therefore do not acquire reputations in the intended sense. When they act so as to project the best possible image of themselves, it is in order to impress conspecifics that are watching them, and not to bolster their reputation with third parties. When we humans try to project the best possible image of ourselves to those around us, it is typically both in order that these observers should form a better *opinion* of us and in order that they should, as a result, contribute to our good *reputation* through future conversations with others.

Opinion and reputation may diverge. Here is a literary illustration. The sister-in-law of Vronsky, Anna Karenina's lover, explains to him that she does not judge Anna, but that she cannot receive her in her home, above all not publicly, since she has to maintain her reputation to protect her daughters' future.

> Don't suppose, please, that I judge her. Never; perhaps in her place I should have done the same. I don't and I can't enter into that,' she said glancing timidly at his gloomy face. 'But one must call things by their names. You want me to go and see her, to ask her here and to rehabilitate her in society; but do understand that I cannot do so. I have daughters growing up, and I must live in the world for my husband's sake' (Tolstoï, 1877/1953, p. 602).

It would be quite possible for most people in a group to individually refrain from judging a person like Anna Karenina as being bad, but nevertheless to anticipate that she will have a bad reputation, to act towards her on the basis of this reputation and even to contribute to spreading it.

In small groups, where everybody knows everybody else, each individual can form an opinion of others on the basis of direct interactions. Nonetheless, reputation has a role to play: by sharing their experiences and judgments, members of a small group can either confirm their opinion of each other, or realise that their opinion is not a generally accepted one. In this latter case, they may reconsider their opinion and align it with the generally accepted one, defend it, or else hide it. In such small groups, individuals also have the means to act directly on the opinions that others have of them, and thus retain a certain degree of control over their reputation.

In the societies and networks that characterise the modern world, where it is common to enter into relationships with people with whom one is not even indirectly acquainted, reputation is often one's only source of information about them at the outset. At the same time, in such networks, it is more difficult to act effectively on one's own reputation. A person's reputation may reach people with

whom she has no direct or even indirect influence; her character and actions are only one of the sources of a reputation that owes a lot to the authority and biases of those who convey it.

In all societies, but particularly in societies where reputation is such a complex phenomenon and plays such an important role, anticipating the reputational effects of one's actions is part of an individual's social competence. In other words, managing one's reputation calls for much more than attending to low-level cues to the presence of others. It calls for the ability to anticipate the reactions of others to one's own actions and attitudes, including their reactions to our reactions, to the actions of third parties, and so on. It calls in other terms for fairly sophisticated strategic thinking. When Vronsky's sister-in-law decides not to receive Anna Karenina, and more generally when people make conscious choices about reputational matters, it is on the basis of mentally entertained reasons rather than of mere 'gut reactions'.

We are suggesting, in other words, that there are (at least) two mental mechanisms involved in the management of one's reputation, an intuitive one illustrated in Haley and Fessler's experiments and a reflective one illustrated in our literary Anna Karenina example. The intuitive one is, so to speak, a kind of 'Machiavellian instinct'. The reflective one involves strategic Machiavellian intelligence (in the sense of Humphrey, 1976, and Byrne and Whiten, 1988). We may imagine situations where the two mechanisms come into conflict. Imagine, for example, that participants in the dictator game are informed in advance of Haley and Fessler's experimental design and past results. In this case, we might suppose that while on the one hand, they would still have a spontaneous tendency to be influenced by low-level cues to the presence of others (such as eyes and noise), on the other hand they would make a reflective effort not to be influenced by these irrelevant cues.

Consider now the reflective management of our *moral* reputation. This involves anticipating the reputational effects of our actions, that is, anticipating how they will be interpreted and commented upon. Behaving only in ways that would secure other people's approval cannot be a systematic policy, since having a good reputation, however important, is far from being our sole objective. Moreover, our reputation is not monolithic: different people we interact with may have different standards, so that there is no way to satisfy them all. When we choose a course of action that may compromise or erode our reputation, at least with some of the people we care about, we can nevertheless try to protect it by providing a favourable interpretation, or in other terms a justification, ourselves. We contribute to our reputation not only by our actions but also by joining in the conversation about them.

As argued by Haidt (2001; see also Mercier, 2011; Mercier and Sperber, 2011), moral reasoning or reflection, as opposed to moral intuition, is used primarily not to guide action but to justify it (and also to articulate our judgments on the actions of others). This does not make moral reflection irrelevant to how we act, but its relevance is indirect. Moral reasoning does not guide action directly, but it may do so indirectly when we end up choosing a course of action because it is easier to

justify. Here is an illustration. Jill, a teacher, feels it would be morally appropriate to give Fred's essay a C. She is generally in favour of some degree of positive discrimination, but she does not feel that Fred, who is a member of a disadvantaged minority although he himself comes from a privileged family, is really entitled to it. Still, she believes it would be easier for her to justify raising Fred's grade than not doing so, and she therefore gives him a B. What we are suggesting is that, while moral intuition is genuinely moral, much moral reasoning is strategic and directed at protecting one's reputation by finding moral justifications for one's behaviour and, if need be, opting for a course of action that can be more easily justified.

Let us speculate a bit further on a theme that would call for a full treatment of its own. A reflective and strategic concern for our reputation leads us to imagine and take into account the way that others might evaluate our action. A possible heuristic consists in adopting an evaluative 'virtue ethics' stance towards one's own moral character, and attributing one's own evaluation to others, modifying it if necessary to take into account the difference between their point of view and one's own. In general, our reputation among people who share our values and with whom we are the most likely to interact is more important than our reputation with people who are morally and practically more distant from us. Furthermore, the greater the divergences between points of view, the higher are the cognitive costs and risks of taking the point of view of others into account. In other words, a strategically justifiable way of defending one's own reputation is to directly seek the respect of none other than oneself, and to align one's moral justifications with one's moral intuitions. Strategic moral reasoning can be used in the service of moral integrity.

While it might be tempting to think of moral reasoning as culminating in a genuine morality—a 'moral consciousness'—controlled by a sense of self-esteem, there is little reason to believe that this is the most common use of moral reasoning. Rather, it is one possible strategy among others. In contrast to this strategy of sticking to genuinely moral choices in anticipation of a global reputational effect, there is the strategy of investing the necessary cognitive resources and aiming, action by action, gesture by gesture, for desired reputational effects. Different socio-cultural contexts, or even different professions—compare, say, politicians and civil servants—can favour one or the other of these two strategies or suggest still other ones. Moreover, the same person may follow different strategies in different social networks, being for instance a genuinely moral person in the family and Machiavellian in business.

6. Conclusion

Being guided by a genuinely moral motivation may be an optimal way to secure a good moral reputation. Still, people are concerned for their reputation and this may be an extra motivating factor in their moral behaviour. Moreover, in both their moral concerns and their reputational ambitions, people deploy reflective competencies, and in particular, the ability to reflect strategically on how other people might react to their actions. Taking all this into account, we have been led

to distinguish two functions of morality, an evaluative one and a behaviour-guiding one, and four psychological dispositions that play an important causal role in guiding moral behaviour and in securing a good moral reputation:

1. A basic, intuitive moral disposition characterised by authentically moral motivations in one's action and authentically moral criteria in evaluating others.
2. An intuitive disposition to attend to cues to situations where one's reputation might be at stake.
3. A reflective disposition to act strategically to defend one's reputation.
4. A reflective disposition to find justifications for moral choices. To the extent that this search is guided by a sense of self-esteem, it may be genuinely moral, but it need not be.

The two intuitive dispositions may be thought of as evolved mental mechanisms, which form part of the mental equipment that allows children to acquire necessary social competencies, and develop through a learning process subject to some cultural variability. The two reflective dispositions, on the other hand, may be better thought of as particular applications of an evolved capacity for reflective thinking directed at influencing others. It would not be surprising to find that these reflective dispositions have a greater degree of both cultural and individual variability.

Our distinction between the evaluative and the behaviour-guiding function of morality helped us understand morality first as a trait that it is advantageous to look for in others and that therefore others look for in us. It is advantageous, then, to appear to be moral and the most effective way to do so is to actually be moral, to conform to the social selection pressure for genuine morality. So, human morality is not intrinsically hypocritical after all. Still, this leaves plenty of room for hypocrisy. As evaluators of others, we want them to be genuinely moral. Our own motivations in acting are anything but simple. Some are genuinely moral; others are self-interested. Among our self-interested motivations is the desire to secure a good moral reputation even when we don't deserve it and often at the cost of hypocrisy.

Dan Sperber
Department of Cognitive Science and Department of Philosophy Central European University
Budapest
and
Institut Jean Nicod, ENS and EHESS
Paris
Nicolas Baumard
Philosophy, Politics and Economics Program at the University of Pennsylvania, Philadelphia
and
Institut Jean Nicod, ENS and EHESS,
Paris

References

Abu-Lughod, L. 1986: *Veiled Sentiments: Honor and Poetry in a Bedouin Society*. Berkeley, CA: University of California Press.

Alexander, R. 1987: *The Biology of Moral Systems*. Hawthorne, NY: A. de Gruyter.

Ambady, N. and Rosenthal, R. 1992: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2), 256–74.

André, J. B. and Baumard, N. 2011: Social opportunities and the evolution of fairness. *Journal of Theoretical Biology*, 289, 128–35.

Andreoni, J. and Petrie, R. 2004: Public goods experiments without confidentiality: a glimpse into fund-raising. *Journal of Public Economics*, 88(7–8), 1605–23.

Barclay, P. 2004: Trustworthiness and competitive altruism can also solve the 'tragedy of the commons'. *Evolution & Human Behavior*, 25(4), 209–20.

Bateson, M., Nettle, D. and Roberts, G. 2006: Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 2(3), 412–4.

Baumard, N. 2010: *Comment nous sommes devenus moraux: une histoire naturelle du bien et du mal*. Paris: Odile Jacob.

Baumard, N., André, J. B. and Sperber, D. forthcoming: A mutualistic approach to morality. *Behavioral and Brain Sciences*.

Bereczkei, T., Birkas, B. and Kerekes, Z. 2007: Public charity offer as a proximate factor of evolved reputation-building strategy: an experimental analysis of a real-life situation. *Evolution and Human Behavior*, 28(4), 277–84.

Bourrat, P., Baumard, N. and McKay, R. 2011: Surveillance cues enhance moral condemnation. *Evolutionary Psychology*, 2(9), 193–9.

Brehm, J. W. and Cole, A. H. 1966: Effect of a favor which reduces freedom. *Journal of Personality and Social Psychology*, 3(4), 420–26.

Brosig, J. 2002: Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior and Organization*, 47(3), 275–290.

Brosnan, S. F. and De Waal, F. B. 2003: Monkeys reject unequal pay. *Nature*, 425, 297–9.

Brown, W. M. 2003: Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evolutionary Psychology*, 1, 42–69.

Bshary, R. and Schäffer, D. 2002: Choosy reef fish select cleaner fish that provide high-quality service. *Animal Behaviour*, 63(3), 557–64.

Burnham, T. C. 2003: Engineering altruism: a theoretical and experimental investigation of anonymity and gift giving. *Journal of Economic Behavior & Organization*, 50, 133–44.

Burnham, T. C. and Hare, B. 2007: Engineering human cooperation. *Human Nature*, 18(2), 88–108.

Butler, J. 1726/2008: *Fifteen Sermons Preached at the Rolls Chapel*. Google eBook.

Byrne, R. and Whiten, A. 1988: *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford: Clarendon Press.

Delton, A. W., Krasnow, M. M., Cosmides, L. and Tooby, J. 2011: Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, 108(32), 13335–40.

Dugatkin, L. 1995: Partner choice, game theory and social behavior. *Journal of Quantitative Anthropology*, 5, 3–14.

Ernest-Jones, M., Nettle, D. and Bateson, M. 2011: Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior*, 32(3), 172–8.

Fehr, E. and Henrich, J. 2003: Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press.

Frank, R. 1988: *Passions Within Reason: The Strategic Role of the Emotions*. New York: Norton.

Frank, R. 2004: *What Price the Moral High Ground?* (Princeton, NJ: Princeton University Press).

Frank, R. 2006: Cooperation through moral commitment. In G. Bock and J. Goode (eds), *Empathy and Fairness: Novartis Foundation Symposium 278*. Chichester: Wiley.

Frank, R., Gilovich, T. and Regan, D. 1993: The evolution of one-shot cooperation: an experiment. *Ethology and Sociobiology*, 14, 247–7.

Gibbard, A. 1990: *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Gintis, H., Bowles, S., Boyd, R. and Fehr, E. 2005: *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MA: MIT Press.

Goffman, E. 1959: *The Presentation of Self in Everyday Life*. New York: Doubleday.

Guerin, B. 1986: Mere presence effects in humans: a review. *Journal of Experimental Social Psychology*, 22, 38–77.

Haidt, J. 2001: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–34.

Haley, K. and Fessler, D. 2005: Nobody' s watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26(3), 245–56.

Hamilton, W. 1964: The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology*, 7, 1–16 and 17–52.

Hardy, C. L. and Van Vugt, M. 2006: Nice guys finish first: the competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, 32(10), 1402–13.

Hirshleifer, J. 1987: On the emotions as guarantors of threats and promises. In J. Dupré (ed.), *The Latest on the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT press.

Hoffman, E., McCabe, K. and Smith, V. 1996: Social distance and other-regarding behavior in dictator games. *The American Economic Review*, 86(3), 653–60.

Hume, D. 1898: *Essays Moral, Political, and Literary*. London: Longmans, Green, and Co.

Humphrey, N. K. 1976: The social function of intellect. In P. P. G. Bateson and R. A. Hinde (eds), *Growing Points in Ethology*. Cambridge: Cambridge University Press.

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 & Emotion*, 17, 429–53.

Krebs, D. 1998: The evolution of moral behavior. In C. Crawford and D. Krebs (eds), *Handbook Of Evolutionary Psychology: Ideas, Issues, and Application*.

Kurzban, R. 2001: The social psychophysics of cooperation: nonverbal communication in a public goods game. *Journal of Nonverbal Behavior*, 25(4), 241–59.

Kurzban, R., DeScioli, P. and O'Brien, E. 2006: Audience effects on moralistic punishment, *Evolution and Human Behavior*, 28(2), 75–84.

Leith, K. P. and Baumeister, R. F. 1998: Empathy, shame, guilt, and narratives of interpersonal conflicts: guilt-prone people are better at perspective taking. *Journal of Personality*, 66, 1–37.

Lewis, D. K. 1969: *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Lewis, H. B. 1971: *Shame and Guilt in Neurosis*. New York: International Universities Press.

Maynard Smith, J. 1964: Group selection and kin selection. *Nature*, 201, 1145–47.

McElreath, R. 2003: Group report: the role of cognition and emotion in cooperation. In P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press.

Mercier, H. 2011: What good is moral reasoning? *Mind & Society*, 10(2), 131–48.

Mercier, H. and Sperber, D. 2009: Intuitive and reflective inferences. In J. Evans and K. Frankish (eds), *In Two Minds: Dual Processes and Beyond*. New York: Oxford University Press.

Mercier, H. and Sperber, D. 2011: Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*. 34, 57–111.

Mifune, N., Hashimoto, H. and Yamagishi, T. 2010: Altruism toward in-group members as a reputation mechanism. *Evolution and Human Behavior*, 31(2), 109–17.

Milinski, M., Semmann, D. and Krambeck, H. J. 2002: Reputation helps solve the 'tragedy of the commons'. *Nature*, 415, 424–6.

Neff, K. 2003: Understanding how universal goals of independence and interdependence are manifested within particular cultural contexts. *Human Development*, 46(5), 312–18.

Nesse, R. 2007: Runaway social selection for displays of partner value and altruism. *Biological Theory*, 2(2), 143–55.

Nesse, R. 2009: Social selection and the origins of culture. In M. Schaller, S. J. Heine, A. in Norenzayan, T. Yamagishi and T. Kameda (eds), *Evolution, Culture, and the Human Mind*. Philadelphia, PA: Lawrence Erlbaum Associates.

Nowak, M. A. and Sigmund, K. 2005: Evolution of indirect reciprocity. *Nature*, 437, 1291–98.

Ohtsubo, Y. and Watanabe, E. 2009: Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior.* 30(2), 114–23.

Piazza, J. and Bering, J. M. 2008a: Concerns about reputation via gossip promote generous allocations in an economic game. *Evolution and Human Behavior*, 29(3), 172–78.

Piazza, J. and Bering, J. M. 2008b: The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, 6, 487–501.

Pradel, J., Euler, H. and Fetchenhauer, D. 2009: Spotting altruistic dictator game players and mingling with them: the elective assortation of classmates. *Evolution and Human Behavior*, 30, 103–13.

Rawls, J. 1971: *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rigdon, M., Ishii, K., Watabe, M. and Kitayama, S. 2009: Minimal social cues in the dictator game. *Journal of Economic Psychology*, 30(3), 358–67.

Ross, L. 1977: The intuitive psychologist and his shortcomings: distortions in the attribution process. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology* (vol. 10). New York: Academic Press.

Schopler, J. and Thompson, V. D. 1968: Role of attribution processes in mediating amount of reciprocity for a favor. *Journal of Personality and Social Psychology*, 10(3), 243–50.

Sheldon, K. M., Sheldon, M. S. and Osbaldiston, R. 2000: Prosocial values and group assortation. *Human Nature*, 11(4), 387–404.

Smith, R., Webster, J., Parrott, W. and Eyre, H. 2002: The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology*, 83, 138–59.

Sober, E. and Wilson, D. S. 1998: *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sperber, D. 1997: Intuitive and reflective beliefs. *Mind & Language*, 12, 67–83.

Sperber, D. and Wilson, D. 1995: *Relevance: Communication and Cognition*. Oxford: Blackwell.

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K. and Okubo, Y. 2004: Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. *Neuroimage*, 23(3), 967–74.

Tangney, J. P., Stuewig, J. and Mashek, D. J. 2007: Moral emotions and moral behavior. *Annual Review of Psychology*, 58, 1–23.

Tinbergen, N. 1952: 'Derived' activities: their causation, biological significance, origin, and emancipation during evolution. *Quarterly Review of Biology*, 27, 1–32.

Tolstoï, L. 1877/1953: *Anna Karenina*. New York: Random House.

Trivers, R. 1971: Evolution of reciprocal altruism. *Quarterly Review of Biology*, 46, 35–57.

Turiel, E. 2002: *The Culture of Morality: Social Development and Social Opposition*. Cambridge: Cambridge University Press.

Van Vugt, M. and Hardy, C. L. 2010: Cooperation for reputation: wasteful contributions as costly signals in public goods. *Group Processes & Intergroup Relations*, 13, 101–11.

Verplaetse, J., Vanneste, S. and Braeckman, J. 2007: You can judge a book by its cover: the sequel. A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, 28(4), 260–71.

de Waal, F. B. M. 1989: *Chimpanzee Politics: Power and Sex Among Apes*. Baltimore, MD: Johns Hopkins University Press.

West-Eberhard, M. 1979: Sexual selection, social competition, and evolution. *Proceedings of the American Philosophical Society*, 123(4), 222–34.

West-Eberhard, M. J. 1983: Sexual selection, social competition, and speciation. *Quarterly Review of Biology*, 58, 155–83.

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

Wilson, D. S. 1975: A theory of group selection. *Proceedings of the National Academy of Sciences of the USA*, 72, 143–6.

Wynne-Edwards, V. C. and Copner, V. 1962: *Animal Dispersion in Relation to Social Behaviour*. Edinburgh: Oliver & Boyd.