Mind Reading, Deception and the Evolution of Kantian Moral Agents

ALEJANDRO ROSAS

If natural selection is how organisms come to be what they are, altruism and morality involve a paradox. Natural selection takes care of weeding out traits that do not have a beneficial effect on the individuals possessing them and this certainly seems to apply to moral agents, who consistently forgo opportunities for self-benefits acquired through the exploitation of others. Hamilton (1964) proved that altruism could evolve towards kin because genes causing altruism towards relatives benefit copies of themselves. Trivers (1971) proposed reciprocal altruism (RA) as a model to explain away this paradox for traits benefiting non-relatives. Notice that the paradox implies either the existence of traits that have been selected for their beneficial effects on others (relatives or non-relatives), or at least the belief in such traits in current biological theory. Unless biologists really claim that the function of such traits is to benefit others, there is no paradox to worry about. Traits that are incidentally beneficial to others involve no paradox, since by definition they have evolved for other reasons. Accordingly, one can solve the paradox either by showing that traits that are usually believed to have been selected for benefiting others have in fact evolved for other reasons, so that altruism is accidental (Connor, 1986); or by showing that the benefit conferred on others rebounds as a self-benefit that dissolves the paradox. I will assume as uncontroversial that this second strategy is the idea behind the concept of RA.1

As a concept of evolutionary biology, RA refers to behaviors as grasped from their functions or selected effects, not from their motives or proximate causes. When Trivers introduced it in 1971, he introduced also the concept of cheating and explained: “Cheating is used throughout this paper solely for convenience to denote the failure to reciprocate; no conscious intent or moral connotation is implied.” (Trivers, 1971, 36) Now, it should be obvious that this explanation applies also to his concepts of altruism and reciprocity. No conscious intent and no moral connotation are implied by them. This is important, because the terms that Trivers chose for his model are taken from the social and moral sciences, where they have those connotations of intention. But if these terms involve no
reference to the motivational aspects of an action, they are not being used in the same sense in which moral philosophers use them. In consequence, the whole enthusiasm about a forthcoming evolutionary explanation of morality based on them is misplaced and involves a category mistake.

Some philosophers view in this way this notorious chapter in the history of interdisciplinary encounters. I think their reaction is understandable; but I also think it is already apparent that evolutionary theory is developing not only to meet this objection, but to deal with problems that have traditionally worried moral philosophers as well. I will briefly review these developments and will then proceed to show how an evolutionary approach can prove useful in dealing with a long-standing philosophical problem concerning the nature of morality.

**EVOLUTIONARY PSYCHOLOGY AND EVOLUTIONARY THEORY**

The philosophical objection poses a challenge to the evolutionary theorist. He could meet the challenge by abandoning the practice of classifying behaviors exclusively on the basis of their effects and accept the need to include, at least in some cases, a reference to the underlying motives or proximate mechanisms. There is no reason why an evolutionary biologist should ignore the question of the evolution of proximate mechanisms of behavior. That is what evolutionary psychology, in a broad sense, is (or should be) about (cf. Heyes & Huber, 2000). For example, it is likely that particular proximate mechanisms evolve to enhance the reliability of behavioral systems in producing their expected beneficial effects. This provides a reason for evolutionary theory to focus on the mechanisms underlying behavior, and even consider the strategy of classifying behaviors on the basis of the various mechanisms producing them, rather than only on the basis of their evolutionary functions or selected effects. Even if the evolutionary function of exchanging benefits with others can be realized both by intentional and by non-intentional behaviors, the first ones perform their function more reliably in changing environments.

Intentions allow organisms to achieve their goals in varying circumstances. In this they are superior to behavioral routines. Necessarily, living organisms have goals that are desirable for them, but it is not true that they always represent both the states of affairs that are desirable and their desirability. It often happens that their goals are just the outputs in normal conditions of wired-in behavioral routines, which are simply triggered by a stimulus presentation. Animals have intentions only when they are able to represent both the states of affairs they desire and the fact that they are desirable. Consider, for example, the tongue-flick mechanism of the frog. It leads to the ingestion of food, but it does not follow that the frog represents the ingestion of food as a desirable state of affairs. Ingesting food could just be the normal output of a behavioral routine automatically triggered by a stimulus. Nothing in the frog’s brain state needs to be equivalent to the judgments:
“That is food” and “I want food”. These judgments would amount to a representation of a state of affairs as something desirable and would enable the frog to judge whether its response to a given stimulus presentation has led to the desired result. The frog would thus be able to change its behavior if, after some environmental change, its response to a familiar stimulus does not fulfill its intention. This is, admittedly, a rough description of the psychological mechanisms that come along with intentions. It is likely that some capacity for causal cognition and means-end reasoning necessarily belong there too. But whatever else one can say about a psychology that involves intentions instead of behavioral routines, one can say that intentions are part of a psychological mechanism that makes for flexibility in the face of environmental change. This is enough to understand their adaptive value.

In the appropriate contexts, the evolution of intentionality will eventually lead to the evolution of intentional communication and of complex mind reading abilities. Very briefly, once organisms have reached a high level of richness and variability in their intentional life, their behavior is no longer controlled by specific routines, but by a mind that adapts to particular circumstances by constantly forming anew intentions to act in one way or the other. If organisms with such minds live in circumstances that select for communicating abilities, what they would likely communicate would usually be dependent on the particular context of the utterance. Therefore, to be able to communicate successfully, these organisms would have to evolve the ability to understand and read each other’s intentions. This would be a complex form of mind reading, namely, not only being able to see other individuals as minds with intentions, but as minds with intentions and beliefs about the contents of other minds. The communicating partners must be able to represent not only each others’ mind, but also that, and how, the mind of each one is represented by the other. This complex psychological ability is considered by an influential philosophical view as a necessary condition and as the fundamental element in the human capacity for language and communication (Grice, 1957; Bennett, 1976).

The origin and evolution of mind reading has already become a focus of interest in evolutionary theory. The evolution of the capacity for representing other minds as minds that mirror each other’s representations is the object of a promising evolutionary hypothesis. It received a clear-cut formulation in Nicholas Humphrey’s well-known 1976 paper: “The Social Function of the Intellect”. The basic thesis is that higher intelligence evolved in a social context, where individual fitness depends on the ability to understand and act according to rules of social interaction that balance competition and tolerance, and to use these rules creatively to promote oneself to the top of the social pyramid. Implicit in Humphrey’s concept of social intelligence is the idea that organisms are able to read each other’s intentions as a form of enhancing the ability to predict each other’s behavior. Being able to anticipate the behavior of competitors was a decisive weapon in the daily struggle for social status. Humphrey claimed that such abilities would
have evolved in species with complex social relationships, such as primates. He claimed this at a time at which other psychologists were trying to gather experimental evidence on a “theory of mind” in chimpanzees, by which they meant precisely the ability to read the intentions of other minds (Premack & Woodruff, 1978).

The attribution of mind reading abilities to primates is controversial (Heyes, 1993, 1998; Whiten, 1996). It is likely that chimpanzees and other primates have partial mind reading abilities that are better expressed in contexts of competition than in contexts of cooperation. This possibility rests on the fact that mind reading abilities in humans have developmental precursors and are decomposable. However, it is hard to formulate this possibility accurately, mainly because developmental psychologists have not yet given a clear picture of the precursors and decomposable elements of such abilities. A recent move in this direction has been made by Hare, Call & Tomasello (2001), discussing an experiment where chimpanzees perform well using information about what others see in a context of competition for food.

Given current theoretical efforts to understand the evolution of mind reading abilities, evolutionary theory will plausibly be able to approach the explanation of moral behavior in the usual sense, namely as a behavior supported by complex intentionality. At the same time, once the ability to mind-read is factored-in among the relevant variables in the evolutionary scenario and counted among the selection pressures that have plausibly shaped our nature, a novel selection mechanism is activated: psychological selection in social interactions. In the remaining part of this paper I want to explore how these developments invite evolutionary perspectives on moral theory that can be empirically fruitful. They can contribute, so I will argue, to the solution of a long-standing debate concerning the nature of morality.

EGOISM, ALTRUISM AND THE NATURE OF MORALITY

The problem about the nature of morality that we will now consider concerns the basic motivation for moral behavior. One of the constitutive elements in moral thought as conceived by some influential traditional theories is this: in thinking morally about what to do, we do not view other persons as means to our ends but view them properly as ends. In other words, we give the needs and interests of other persons the same consideration and weight that we give our own needs and interests. And we do this conditionally, on the assumption that they will treat us in the same way. Let us call this view of moral persons and their motivations the high-view of morality. The high-view is not a universal feature of moral theories. Some versions of contractarian thought, notoriously Hobbes’ moral theory, advocate a diametrically opposed view of humans as egoistic individuals complying with morality out of a selfish calculus. Call this the low-view of morality. The low-view is as old as Plato’s Republic. One of the characters in that
dialogue, Glaucon, argues for a view of morality that is very similar to the one developed by Hobbes in modern times. In a nutshell, it is the view that morality is a human invention, a device required for the satisfaction of individual needs when individuals cannot coerce or manipulate others to their own purposes and lack the self-sufficiency to avoid interaction with them. In this case, individuals must cooperate with each other for mutual benefit, but they do not do so because they enjoy cooperation or express through it their respect of others’ interests. Their cooperative attitude is only instrumental and it is maintained insofar as it serves selfish purposes. Some moral philosophers argue that the low-view cannot explain our everyday moral experience; but for Hobbesians this only means that our everyday beliefs about morality are nothing but an illusion.

Interestingly, similar doubts about the high-view of morality can be fuelled from a biological perspective. The fundamentals for this line of thought have been laid down by evolutionary biologists. It was first suggested by T. Huxley and his idea of an unavoidable conflict between ethics and evolution, but it has experienced a rebirth among contemporary evolutionary theorists that insist on looking at evolution from the perspective of the gene (selfish-gene-theory) (Dawkins, 1976). From this perspective organisms will have, directly other indirectly, their reproduction as their ultimate goal because of the way natural selection works. In organisms with minds and intentions this goal will presumably be represented intentionally, though perhaps unconsciously, as the ultimate motive guiding all action (Alexander, 1987). Such organisms will be psychological egoists, and will confine their altruistic motivations, if they have any, to offspring and near relatives. Support for this view can also be drawn from evolutionary theory’s newborn interests in the evolution of mind reading. As reported above, a promising hypothesis states that mind reading evolved in the context of competition. The ability to read the intentions of others and anticipate their behavior served the individual pursuit of social status and the struggle to increase one’s share of mates and resources. In short, it evolved as an ability to better use others to one’s selfish ends. If the evolution of mind reading abilities is an important fragment in the evolutionary explanation of morality, morality seems to be something in line with what the low-view has suggested.

One could object that mind reading abilities are also useful in cooperative and moral (high view) contexts. High view morality holds, characteristically, that in thinking morally about what to do we do not view others as means to our ends but as ends themselves and we give them the same consideration and weight that we give our own person. And we do this conditionally, on the assumption that they will treat us in the same way. This is an altruism of sorts, or more properly, fairness. Obviously, morality as conceived in the high view requires mind reading as well. Each agent knows that cooperation from others is conditional upon his behavior and he attributes others the same knowledge regarding the conditional nature of his cooperation. To behave morally towards each other, we have to
conceive of each other as a mind with intentions to act in one way or the other. In consequence, it is theoretically possible for mind reading to evolve as subservient to a moral (high-view) way of thinking.

This line of thought has led some authors to object that selfish-gene theory too quickly infers psychological selfishness from evolutionary selfishness, confusing evolutionary functions with proximate mechanisms. Frans de Waal (1996) views this confusion as the mark of “calvinist sociobiology”. Against “calvinist sociobiology”, it is certainly possible for evolutionary selfishness to promote psychological altruism, and it has been argued that this must be the case in parental care (Sober & Wilson, 1998). A parent caring for her offspring on the ground of her own pleasure would be less reliable than one caring for them for their own sake. Therefore, natural selection operating on individuals would favor altruistic parents, as those that on average raise more children to reproductive age. But note that this criticism of selfish-gene theory does not bite through. Selfish-gene theory does not need to argue that humans are radical psychological egoists. It can admit psychological altruism in humans, provided that it is confined to offspring and near relatives. This would mean that we are biologically designed for nepotism, and that we do not develop altruistic feelings towards strangers in large societies, as the high-view of morality requires. Alternatively, some have tried to explain high-view morality as a sort of “spandrel”, a by-product of other adaptations, but this view has not been yet thoroughly worked out (for the best try see Singer, 1983).

The view that RA, cooperative behaviors and the mind reading abilities required for them have co-evolved with high-view morality should, I believe, be taken seriously. But unless we find some convincing reasons for favoring high-view morality and its motivational underpinnings, it remains open for “calvinist sociobiologists” to interpret RA in humans as resting on psychological egoism, in line with the selfishness presumably underlying the evolution of mind reading. Such an interpretation receives support from the philosophical perspective pursued by the Greek sophists, by Hobbes and by other contractarian writers in modern times, who have noted that cooperation between individuals will emerge under specific circumstances, notably when resources are scarce or otherwise not retrievable by individuals acting alone; when individuals competing for them are predominantly interested in their own needs; and when it is not possible to coerce or maintain coercion over others in the long term. These conditions are similar to those that Trivers held would favor the evolution of RA. Cooperation and RA can thus rest on egoistic motives. This is the strength of the low-view of morality, namely, its ability to explain cooperation and the practices of benefit-exchange as resulting from a mechanism of mind reading sub serving selfish motives. Therefore, we cannot just conclude that the need to cooperate must have resulted in the selection of altruistic motives in our ancestors.

However, the low-view of morality seems to evoke in many of us feelings of rejection or even disgust that provide at least some evidence against this view. The
reason for these feelings is simply brought out by mentioning what the view lacks: it lacks the idea that we should meet other persons with the belief or the feeling, or both, that they are not merely means to our selfish ends, but ends deserving respect as such. If the low view of morality were true and its truth were acknowledged, we would have to live with the thought and feeling that each and everyone of us cooperates with others only when the prospects of cheating and getting away with it are weak, while silently awaiting for the opportunities where those prospects increase and a ruthless pursuit of individual advantage promises to pay. I assume that this is not a world where most of us would feel at home, but I have not found in contractarian theorists an explanation of why we feel this way. If morality is what the low-view says it is, why do most of us find this thought so troubling?

Perhaps surprisingly, new impulses to think about this problem have come from evolutionary biologists that developed their ideas in an interdisciplinary framework embracing psychology and moral philosophy. I am thinking specifically of R. Trivers and R. Alexander. In what follows, I will borrow freely from them and from work done by moral philosophers sympathetic to evolutionary approaches, especially work done by D. Gauthier and his critics. My purpose is to develop briefly and clearly a plausible evolutionary argument for the thesis that our evolved moral psychology is essentially of a Kantian type.

EXPLAINING OUR MORAL FEELINGS: SELF-DECEPTION

R. Alexander’s *Biology of Moral Systems* deserves special attention in this context. Its significance rests, in my opinion, on the central importance he attributes to explaining why our everyday feelings are against the low-view of morality and broadly agree to a Kantian conception of moral psychology. Since Alexander endorses the low-view, he has proposed an explanation that blocks the attempt to use these feelings as evidence against it. His explanation is embedded in an evolutionary theory of moral systems based on a development of Trivers’ model of reciprocal altruism. I offer here what I see as his core argument concerning this question.

Endorsing selfish-gene theory and the idea that biological organisms have reproductive success as their foremost interest, Alexander believes that humans have evolved to be psychological egoists, although he allows for the existence of altruistic motives towards offspring and near relatives. It is within this basic egoistic/nepotistic framework that he endorses Trivers’ concept of reciprocal altruism, emphasizing that in human societies systems of reciprocity are generalized or indirect (Alexander, 1987, 93–103). In human societies everybody keeps an eye on the behavior of third parties as a means of acquiring knowledge that can be useful for future interactions. Though human agents are egoists, the awareness of their behavior being generally monitored leads them to be concerned about building a deceptive reputation as altruists. The argument assumes in its critical
turn that this is the best way for egoists to take advantage of the opportunities for cooperation.

To understand the argument one has to place it in the context of the evolution of mind reading. As humans (or hominids) evolved the ability to read the intentions of others as well as their own, they came to understand the unavoidable selfishness driving human action. This was an obstacle for building cooperative partnerships among non-relatives, because individuals prefer to interact with Kantian altruists rather than with egoists. Alexander assumes this as a fact about human preferences on the basis of common sense. He does not give any empirical or experimental support. I believe his assumption is basically correct, and will refer later to provisional experimental support. Therefore, in Alexander’s evolutionary scenario humans faced a problem that natural selection had to solve for them: while each of them needed partners for cooperation, the selfish nature of basic human motivations seemed to preclude the possibility of attracting partners. Their only chance consisted in being able to conceal their selfish motivations; and since conscious deceivers too often betray themselves involuntarily, the best way to do this is by concealing those motivations from themselves (Alexander, 1987, 123). Thus, natural selection favored the evolution of self-deception. We conceal our selfish motivations from ourselves only to better conceal them from others. A fundamental element in this self-concealment is the fact that we denigrate selfishness and praise altruism. Our conscious values are designed to maintain deception and self-deception. If we are shocked and disgusted at theorists defending the low-view of morality, this is just a product of self-deception, and should not be used as evidence against the truth of the low-view.

Alexander defuses in this way the testimony of everyday moral thought against the low-view of morality. But closely examined, his explanation of our moral feelings is circular. In the scenario for the evolution of self-deception informally envisaged here, the starting point of the process consists in non-relatives meeting each other with mistrust in cooperative interactions, given their ability to read each other’s inevitable selfish motives. But this mistrust is precisely an aspect of the same phenomena that the theory should explain. Mistrust of selfish cooperators and preference for Kantian partners are expressions of a value system that praises altruism and denigrates selfishness. As we saw above, Alexander views this value system as subsidiary to self-deception and the deception of others. So Alexander is trying to explain these attitudes through self-deception, but he also explains the evolution of self-deception by picturing a scenario where humans feel mistrust towards selfish partners and know that others feel in the same way from the start. Only because these attitudes were already there at the beginning, a selection pressure existed for the evolution of self-deception in agents desperately needing to attract partners for cooperation and aware of the obstacle built by their—unavoidable, according to Alexander—selfish motives. As an explanation of self-deception this is plausible, but it requires the previous existence of a negative valuation of selfishness. If dislike of selfishness in potential partners was there
before self-deception evolved, then this dislike demands an explanation that does not appeal to self-deception. We cannot logically explain mistrust on the basis of something that is in turn explained by the previous existence of mistrust. We must therefore look for an alternative explanation of our mistrust of selfish partners or of the value we attach in our everyday thought to moral agents of a Kantian type. I offer one below that has no need to question as self-deceptive our belief in the reality of fairness and altruism.

**DISPOSITION FOR FAIRNESS AS SIGNAL OF RELIABLE COOPERATORS**

I will now sketch an evolutionary scenario that explains the evolution of a Kantian moral psychology as a fact, rather than as a self-deceptive belief, about ourselves. The scenario is constructed to identify a particular selection pressure that accounts for the value we attribute to altruism as fairness in our everyday moral thought. I will borrow from Alexander’s model the idea that the capacity to deceive played an important role in this process. In fact, deception is tied to mind reading in this way: if we were perfect mind readers, deceiving others would not be an option. It is and was an option only because we are imperfect mind readers. Let us label “transparency” the dimension of social interaction that depends on the accuracy of mind reading. To see how the degree of transparency affects the evolutionary scenario, consider first what happens when our scenario is populated with rational agents who meet each other to interact in a condition of perfect transparency. We assume that this scenario is governed by the conditions that Trivers and the contractarians mention as favoring cooperation and reciprocal altruism. In a condition of perfect transparency, every agent knows the intentions and the behavioral strategy of its potential partners as well as it knows its own. If the scenario is populated only with selfish agents, who place their needs and interests above anybody else’s, transparency dictates that cooperation will be the best mutual move, since any intention to defect will be transparent and will cause one’s partners to defect as well. If selfish agents do what is best for them, they will be willing to cooperate with each other on a reciprocal basis in order to reap the benefits of cooperation. They will follow the rule: “I will cooperate with you if you cooperate with me and I will not if you do not”. Under the condition of perfect transparency, selfish dispositions ensure cooperation and have the virtue of building, as it were, a perfect illusion of morality. Cooperation flourishes in this world. Even if one insists that this is only an illusion of morality and not the thing itself, the thought experiment arguably shows that egoism by itself is not the principal target of our negative feelings towards the low-view of morality. In this world anyway, co-operating agents are obviously not bothered by the selfishness of their partners, as long as they are willing to cooperate on a reciprocal basis.

Enter now Kantian agents. These are essentially fair players: they place their needs and interests on an equal footing with those of others. They will not
cooperate with those intending to defect on them, since this would place their own needs below those of defectors, whereas Kantians give equal consideration to everyone’s needs. So they will display the same conditional cooperative strategy as selfish agents. If they are not bothered by the egoistic motives of their selfish neighbors, they will cooperate with them as long as these are willing to cooperate too. Given transparency of intentions, Kantian agents cannot be harmed by selfish ones. But suppose Kantians are hard-liners, and as such are in fact bothered by the motives of selfish agents. Hard-liners require from their partners not only cooperative behavior, but also that they cooperate out of fair motives (out of a respect for persons as ends and not only as means). In this case Kantian agents will not cooperate with selfish agents. And because of the transparency condition, selfish agents will know this and will deny them their cooperation in turn. If these are the only types of agent present, it is intuitively clear that none will have an evolutionary advantage over the other. Those of the same motivational type will cooperate among themselves and deny cooperation to those of the other type. Egoists will do as well among themselves, as altruists will do among their kind. Motivational altruism is not universally valued in this world. In fact, if cooperation is not universal, only Kantian agents and their “intolerance” towards selfishness are to blame.

Let us now change the condition of transparency. We abandon the assumption of perfect transparency and assume partial opacity. The condition of partial opacity will prove decisive for the evolution of morality as conceived by the high view. As long as transparency is less than perfect, mistakes in judging others become possible. More importantly, some mistakes will be deliberately induced by others; this is what we call deception. Assuming that Kantians will resist the temptation of deceiving others, only selfish agents will employ deception as a strategy. Since selfish agents put their needs above those of everybody else, their rational calculations concerning the best strategy to pursue will look different in a scenario where perfect transparency has given way to partial opacity. They will predictably consider the chances of gaining greater payoffs through the exercise of deception. Under the condition of partial opacity, selfish agents introduce something that could not exist before, namely cheating and the exploitation of others. Though selfish agents will be the only perpetrators of these strategies, everybody is a potential victim. So everybody will be concerned with securing cooperation and avoiding exploitation; and the best way to do this is to choose or discard partners on the basis of signals of their motivational dispositions. Under the condition of partial opacity Kantian agents and their motivational disposition for fairness become a precious good in the eyes of everybody. This is so because fairness of character is the only reliable sign of non-deceptive strategists in a world where transparency is less than perfect. And since cooperation is actually the most reliable way to most goods (both in our thought experiment as well as in the actual world) we value altruism and fairness highly in our value system. The concern with reputation, one’s own and that of others, is probably a consequence of partial opacity.
By picturing an evolutionary scenario under variation of the condition of transparency, or of the accuracy of our mind reading abilities, we illuminate the fact that those abilities, which make human cooperation and morality possible, have exerted an important selection pressure in the evolution of those feelings that are so central a part of our moral attitudes: the feelings aroused by the motivational dispositions and characters of others. The contrasted scenarios help us understand the grain of truth contained in the low-view of morality. In a world of perfect transparency and rational actors, selfishness can be a sign of reliable cooperation. It is conceivable that we would not have to feel disgust towards egoism in such a world. It is only because we are imperfect mind readers that egoism becomes the vehicle for nasty strategies of social interaction and has evolved to be the target of aversive feelings. This happens because deception is incorporated into the strategies of egoists. When deception plays a role, a motivational disposition for fairness is highly valued in partners, because it is the only reliable sign against the use of deception.

Our model assumes that our mind reading abilities, though imperfect, are accurate enough to support a strong psychological and social selection in favor of Kantian agents. Although the model is theoretical, it pretends to explain our real psychological preferences. That we do in fact prefer altruistic to selfish partners in cooperative interactions (especially those with a PD structure) is a thesis that, though it still awaits full experimental confirmation, is provisionally suggested by recent experimental reports. In particular, Sheldon et al. (2000) reported an experiment in which participants that were free to choose partners in a Prisoner’s Dilemma game with a real resource at stake, preferred to choose partners with a strong pro-social orientation.

The model preserves one of the insights that makes Trivers’ model different and superior, in my opinion, to the one conceived by Alexander. It is the idea that a psychological selection for altruistic or fair motives in social agents has taken place, because partners in a cooperative enterprise are judged by their motivational dispositions, which are the only reliable signs of a cooperative attitude (Trivers, 1971, pp. 50–51). I have here developed this idea by completing our picture of the selection pressures involved in the evolution of the preference for fair players. Being imperfect mind readers, we cannot place our trust on the rational egoism of our partners, but only on the fact, where it is a fact, that they have strong dispositions towards altruism as fairness.

Alejandro Rosas
Departamento de Filosofía
Universidad Nacional de Colombia
Ciudad Universitaria
Bogota
Colombia
arosasl@unal.edu.co
Acknowledgements. I would like to thank the Alexander von Humboldt Foundation for financing the research activities that made this paper possible. Thanks are also due to the Universidad Nacional de Colombia and to Prof. Ludwig Siep at Münster University (Germany) for their institutional support.

NOTES

1 Since Trivers’ work on RA, the fitness and adaptiveness of such a behavior relative to other possible strategies has been thoroughly explored and confirmed with the aid of game theoretical arguments, starting with Axelrod & Hamilton (1981).

2 I am using one of Kant’s formulae for the categorical imperative as representative for this view of morality. But I expect the formula to stand on its own in this context. Besides, the added conditional clause cannot claim his authority.

3 Alexander in fact says that people prefer to interact with “honorable egoists”. But in his usage, an honorable egoist is someone who has internalized moral rules and follows them because they also serve his ends (1987, 118). I believe this usage lags behind recent work on the distinction between egoism and altruism. To preserve the distinction, I here use “egoist” to denote those people who act always to serve their own ends and who follow moral rules, if at all, not because they also serve their ends, but only because and only when they serve their ends. Cf. the work done by Dan Batson and colleagues on the empathy-altruism hypothesis. Cf. Sober & Wilson (1998 ch. 7) and my discussion (Rosas, 2002).

REFERENCES

Hare, Call & Tomasello (2001). Do Chimpanzees know what their conspecifics know? Animal Behaviour, 61, 139–151.


