Chapter 2
Rationality and Deceit: Why Rational Egoism Cannot Make Us Moral

Alejandro Rosas

2.1 Human Cooperation and Evolutionary Altruism

Cooperation is a pervasive phenomenon in the biological world. Evolutionary biologists hold it responsible for the existence of hierarchical levels of biological organization. Genomes and multi-cellular organisms behave as individuals and are treated by scientists as such, but they evolved out of independent, lower level biological units, in a process e.g. from single cells to groups of such cells, before those groups evolved into multi-cellular organisms as individuals in their own right. Biologists believe with good evidence that cooperation between lower level units drove this process (Buss 1987; Maynard Smith and Szathmáry 1997; Michod 2007). Through cooperation, individuals at the lower level obtained benefits that were not otherwise available. In this process, either cheating was not a problem, or natural selection had to solve it. It is not exaggerated to say that evolutionary biologists place nowadays as much emphasis on the role of cooperation in evolution as traditionally was placed on competition.

The ubiquity of cooperation in the biological world suggests that cooperation among humans has a biological basis. The claim is that we have been designed by natural selection to cooperate throughout history and pre-history at the large scale peculiar to humans. In this paper I shall not present or develop arguments explicitly defending natural selection as the designer of humans as co-operators. Those interested in the arguments for an evolutionary explanation of human cooperation can find them in the first four chapters of Joyce (2006). (I assume, as Joyce does, that cooperation is the main theme of moral norms.) Most evolutionary biologists assume the legitimacy of the evolutionary approach to morality. Their efforts in this subject have been mainly devoted to solve a particular evolutionary problem.

A. Rosas (✉)
Philosophy Department, Universidad Nacional de Colombia, Bogotá, Colombia
e-mail: arosasl@unal.edu.co

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_2,

Morality seems to involve a paradoxical lifetime sacrifice in fitness to the benefit of others, including those that are the least moral in the group. Why doesn't selection eliminate moral traits, if it selects the fittest organisms? Biologists have at least three theories available to solve this paradox: kin selection, reciprocal altruism and group selection. Group selection enjoys popularity with some philosophers and in particular with a group of social scientists, who advocate strong reciprocity to explain human cooperation, a strategy that requires group selection to evolve (Sober and Wilson 1999; Gintis et al. 2003).[1] Darwin was perhaps the first scientist to advance a group-selection hypothesis. He was painfully aware of the paradox involved in viewing human morality as an adaptation and speculated that groups of morally motivated humans proved more adaptive than groups of selfish individuals in ancestral intertribal warfare (Darwin 1981, chap. 5).

For those endorsing an explanation based on group selection, human cooperation is biologically altruistic. This means, as stated above, that the cooperative agent suffers a paradoxical lifetime sacrifice in fitness to the benefit of other individuals. The sacrifice is paradoxical because, *prima facie*, a trait that sacrifices fitness should not evolve. Since there are currently rival proposals about how the sacrifice is to be conceptualized and measured, the concept of evolutionary altruism is bedevilled with uncertainties that are getting more complicated as the debate develops. For example, most evolutionary theorists consider reciprocal altruism as a biologically selfish (fitness enhancing) trait (Lehmann and Keller 2006; West et al. 2007): it entails a short-term sacrifice in fitness, but in the long term the trait enhances the fitness of its carrier. Nonetheless, reciprocal altruism is potentially vulnerable to exploitation, which was probably the reason why Trivers used the label 'altruism'. In any case, if a helping trait is to evolve, it must have higher fitness than rival traits: labelling a helping trait 'altruistic' in the sense of involving a fitness sacrifice apparently contradicts this obvious fact. Biologists invoke differences between direct and indirect fitness effects (West et al. 2007), or between within-group and between-group fitness (Sober and Wilson 1999) to legitimate the view that some traits really sacrifice the fitness of their carriers and, nevertheless, are selected. Except for by-product mutualisms, which clearly involve no sacrifice (Sachs et al. 2004), cooperative traits seem both to enhance and to sacrifice the fitness of carriers, depending on the point of view.

The question of evolutionary altruism is a tricky one. I gladly endorse a conceptual reform based on the concept of positive assortment, as suggested in some recent literature (Fletcher and Doebeli 2009; Rosas 2010; Bowles and Gintis 2011, table 4.1, p. 75). But luckily, I do not need to discuss this here. For the purposes of this paper, I am happy to follow Joyce (2006, pp. 16, 38) in adopting an uncommitted view on the question of evolutionary altruism. Natural selection favours helping or

[1] For a recent sophisticated defence of strong reciprocity and group selection see Bowles and Gintis (2011), especially chapters 6 and 7. For a review see Rosas (2012).

cooperative traits and, for the purposes of this paper, I leave open whether these traits enhance the fitness of their carriers and evolve for this reason, or sacrifice it and evolve for benefits they receive in an indirect, though reliable, manner.

2.2 Social Preferences Versus Selfish Cooperation

The question addressed in this paper is, namely, whether social preferences, specifically altruistic ones, are among the proximate mechanisms explaining cooperative behaviour and morality in humans; and indeed, whether they are indispensable. Social preferences, as in general other conative mental states (motives, desires or emotions), are altruistic if they satisfy two conditions: (1) their content is the welfare of another person or persons, and (2) wanting another's welfare is not instrumental to the agent's own welfare, such that the agent would stop wanting it if she could promote her welfare by other means. An altruistic preference is therefore a non-instrumental positive interest in the welfare of some other person or persons.

When saying that social preferences are necessary both for morality and for cooperative/moral behaviour, I do not mean to conflate morality with behaviour that is merely in accordance with moral demands and was not prompted by motivations that we would deem moral. In particular, I do not want to belittle the importance of judgments with the predicate 'is wrong' as necessary elements in motivations and dispositions that we consider moral. Joyce (2006) has argued this point cogently. However, these considerations are compatible with the view that social preferences, though not sufficient, are necessary for moral judgment. We would not be able to make moral judgments without them. The judgment 'X is wrong' arises, on this view, only if the subject experiences a conflict between the opposing requirements of selfishness on one hand, and a non-instrumental (altruistic) inclination or desire that others do well on the other. This conflict emerges particularly in situations known as social dilemmas. In these situations, it is wrong to follow the morally selfish temptation of profiting at the expense of others' contributions.

Positing social preferences as necessary conditions for cooperation challenges the views of classical economists and game theorists. Their model of the human agent, *homo economicus*, is a non-tuistic agent, i.e., an agent who lacks social preferences. Nonetheless, classical economists and game theorists expect cooperative behaviour from *homo economicus* in repeated games, because cooperation is what rational egoism demands. Recently, this view of cooperation based on an 'invisible hand' coordinating the desires of selfish agents has been challenged by social scientists pointing to the 'dark side of self-interest' (Bowles and Gintis 2011, pp. 5–6). They disapprove of *homo economicus* as a complete explanatory model and demand the inclusion of social preferences, in particular altruistic ones, in the model of a human agent.

Game theorists and classical economists believe that humans cooperate on the basis of egoism because they think selfish rational agents can reason through to the conclusion that cooperation is in their best interest, particularly in iterated

prisoner's dilemma games (iPD). Rational egoists run through a normative argument to the effect that they ought to cooperate in iPD games to enhance their own material benefits. On this ground, rational egoism appears as a sufficient explanation for cooperation. One version of the normative argument is the folk theorem in game theory, proving the existence of (many) Nash equilibriums with cooperative strategies in iPD games. Therefore, a rational agent ought to choose one or another cooperative strategy. Another version of the argument is that of Gauthier who tries to prove that even in one-shot PD games cooperation is rational (the best response) under conditions of transparency or translucency (Gauthier 1986).

For those who believe in some role for social preferences in human cooperation, the question arises: what exactly is their role in human cooperation? If rational egoists, (rational agents lacking social preferences) ought to cooperate in their best interest and if they know this, why do they have social preferences and do they (we) really need them? If we need them, shouldn't we simply mistrust the normative argument based on rational egoism? It will be useful to provide here the basic tenets of three different answers to these questions.

Many authors grant that actual cooperative behaviour is more likely based on moral emotions than on a rational argument. Nonetheless, the normative argument need not be superfluous. It offers a justification of the rationality of social and moral emotions. Gauthier, for example, acknowledges that moral behaviour is usually driven by moral emotions (instantiating social preferences like fairness), but the normative question is still worth pursuing, for it examines whether the behaviour prompted by those emotions is rationally justified or should be discarded as irrational (Gauthier 1986, p. 338). His theory thus offers a 'rational reconstruction', and gives reasons for being moral that are not the reasons humans actually follow. But supposing some humans are really able to act on the reasons provided by the normative argument, it does not follow that social emotions or preferences are superfluous. Minds can be designed with redundant mechanisms for purposes that are important for survival, as cooperation in this case.

However, alternatively, it may be the case that humans, being imperfectly rational, are not able to act on the reasons provided by the normative argument unless social emotions come to our aid. Imperfect rationality is the basis for another explanation of the role played by social preferences. They are not simply backup mechanisms: we have them because we need them. They help us comply with what is rational. In this explanation, cooperation is rational for agents that lack social preferences; and perfectly rational agents would be able comply with this rational demand even in the absence of social preferences. But humans, being imperfectly rational, cannot comply because, e.g., of excessive temporal discounting. Social preferences are then required to remedy our imperfect rationality and bring our behaviour back in line with what is rational, namely cooperation. Frank (1988) has pursued this sort of argument, and Joyce (2006, chap. 4) follows this same line.

Finally, a still different explanation says that cooperation is not the rational move for rational agents that lack social preferences. Rationality will not lead rational agents to cooperate with each other if they lack social preferences, because they foresee that their payoffs will be higher if they can coerce or deceive others. In the

following I shall argue for this third view in an evolutionary setting. Cooperation would collapse in a human population in the absence of a sufficient proportion of agents with social preferences. And since cooperation is important for survival, we are lucky that natural selection managed to graft social preferences onto our rational mind.

2.3 Selfishness and Deceit

I have sketched three possible evolutionary explanations for why we have social preferences as evolved proximate mechanisms for cooperation. I assume that we do have social preferences. Since perhaps not everybody is willing to grant this, I mention that evidence in favour has been accumulating lately in the field of experimental economics (Bowles and Gintis 2011, chap. 3). With these two assumptions, I now concentrate on arguing for the third view. The third view says that rationality does not recommend cooperation as the best move if rational agents lack social preferences, i.e., if they have no interests in the interests of others or are mutually unconcerned (the interests of others do not figure in their utility function). The reason lies in our imperfect mind-reading abilities. Having imperfect knowledge of the practical intentions of other agents, deception and several forms of coercion are rational for agents without social preferences. Mary Gibson (1977) and Jean Hampton (1991) have made this sort of argument before. I here place it in an evolutionary context.

Assuming natural selection designed the human mind with what it takes to make cooperation possible, there is a wide consensus identifying the basic selection pressure for the emergence of cooperative traits in humans: it is a general structure of interaction represented in the iterated prisoner's dilemma (iPD). In a simple one-shot PD, cooperation is rational and adaptive only if opponents are conditional co-operators that predict reasonably well the intention of their co-players. In repeated games, where players A and B will interact many times, player A can punish through defection at $t + l$ a defection by B at period t. In this case cooperation is also the best response. Punishment is crucial for the stability of cooperation in iPDs. It encourages would-be defectors to cooperate in order to profit from the opportunities of mutual gain (Trivers 1971; Axelrod and Hamilton 1981).

However, repeated interaction is not sufficient and some extra conditions are required. For example, players must have equal power. A dominance relationship prevents the emergence or stability of cooperation, if cooperation is conceived as instantiating a principle of fair distribution of profits (after subtracting costs). An unfair distribution is not cooperation, even if both players gain something beyond their investments. But suppose interaction is free from coercion; cooperation could still fail to evolve. If the iPD is finite, as it always is, there is a temptation to lure others into cooperation only to cheat grandly on the last round. An experienced cheater plans the last round beforehand and then disappears. In this case, cooperation preceding the last cheating move is selfishly instrumental

to the substantial profits obtained in that move. This strategy poses a problem for cooperation. It corresponds in spirit to Hobbes's Fool and Hume's Sensible Knave. Those characters use cooperation instrumentally and selfishly to maintain a deceptive reputation for honesty and to create opportunities for exploitation. They endorse morality and the cultivation of a good reputation not for their own sake, but instrumentally: they behave fairly or honestly only when not doing so would damage their reputation. Notwithstanding honesty being a good general policy, it is subject to many exceptions, so that the (egoistically) wisest man is he who observes the general rule and free rides on all the exceptions (Hume 1902, §232). The instrumental cultivation of a reputation for being moral is essentially deceptive (Sayre-McCord 1991). Deceivers will do whatever it takes to appear as moral individuals in the eyes of others, but whenever they can cheat undetected, they will.

The combination of selfishness and deceit is perhaps the crucial objection against the view that cooperation is rational for agents without social preferences. Despite Hume's official doctrine that justice arises from rational self-interest, his words about the Sensible Knave seem to confirm that rationality is not what he lacks, but rather a primitive motive for justice: 'If his heart rebel not against such pernicious maxims, if he feel no reluctance to the thoughts of villainy or baseness, he has indeed lost a considerable motive to virtue' (Hume 1902, §233). In contrast, Hobbes overtly argued that the Fool is in fact irrational (Hobbes 1651). But Hobbes's argument works only if you assume that the Fool cannot deceive others. If he can, he is rational to profit at their expense. Therefore, morality cannot rest merely on selfish motives and rational choice. Morality requires a primitive disposition to take persons as equals, a quality that cannot be constructed out of the rational choice of non-tuistic agents (Hampton 1991). This disposition blocks the use of deception as a rational option, because deceiving others contradicts a sense of fairness. Deceit is a rational option for agents who lack social preferences, but only for them. In this sense, the Hobbesian, rational choice derivation of morality is mistaken. Morality cannot be derived from rationality if rational agents lack social preferences. Human cooperation rests necessarily on an irreducible and non-instrumental motive for fairness.

2.4 A Theory of Morality as Disguised Selfishness

The view that human cooperation is based on rational egoism appeared originally among philosophers, namely among the Greek Sophists as depicted by Plato in *The Republic*; and then in the *Leviathan* of Thomas Hobbes. It has been taken up by some evolutionary theorists of social behaviour, of whom I shall mention Richard Alexander (1987), who plays an important role in the argument to follow. It is important to confront and refute this view, if social preferences are to be advanced as evolved mechanisms necessary for cooperation, and not simply, if at all, as backup mechanisms to rational choice. Joyce (2006, p. 17) criticizes Alexander's view, but misrepresents it as a naïve conceptual confusion. Regarding Alexander's claim that

selfishness is the motive for cooperation in humans, Joyce says: 'But such attitudes, posing as hard-nosed realism, erroneously conflate distinct explanatory levels (see Tinbergen 1963). In particular, they commit the basic blunder of confusing the cause of a mental state with its content.' (Joyce 2006, p. 17).

Joyce claims, as Tinbergen (1963) could have put it, that Alexander conflates the evolutionary with the proximate cause: if the social trait evolves because it brings fitness benefits to genes or individuals (evolutionary cause), then the mechanism driving it, in case it is an intention, *must be* the intention to produce a benefit to the individual or its genes, a selfish intention (proximate cause). But in fact, Alexander is not guilty of such naïve confusion. He is perfectly aware that selfish genes can produce psychologically altruistic motivations, for example between close kin (Sober and Wilson 1999, chap. 10, have spelled out an argument that shows how 'selfish' genes code for psychological altruism in parental care, namely because motivational altruism causes parental care more reliably than egoism, and parental care is crucial for reproductive success in humans). Alexander knows that selfish genes *do not have to be* expressed in selfish motivations in every case. He does think that they *are* so expressed in the relations between non-kin, but he thinks this on the basis of an argument. His reasons for believing that selfishness governs interactions between non-kin in large societies tally with those that convinced Hobbes, or the Smith of *The Wealth of Nations*. We cannot hold Smith or Hobbes guilty of a conceptual confusion involving proximate and evolutionary causes. In fact, Alexander's theory is probably the strongest objection to the view that genuine social preferences are necessary for human cooperation. His theory is designed to explain away the philosophical claim that moral systems rest on social preferences or non-instrumental desires for the welfare of others. According to him, cooperation is possible among selfish agents making widespread use of deception and self-deception in their mutual interactions. This theory is perhaps his main reason for believing that our motivations are mainly selfish.

Alexander's view relates neatly to the two characters already introduced: the Fool and the Sensible Knave. The core of these characters is to endorse morality and honesty only instrumentally. This was, for Hume at least, a distortion of the place that morality really occupies in the human mind, although he was prepared to accept the Knave as a rational character. In contrast to Hume, Hobbes believed that morality is instrumentally rational for selfish agents and that it is not rational to take advantage of others by deceiving or coercing them (Hobbes 1651). The end of the last section casts a doubt on Hobbes's claim that the Fool is irrational. In any case, the claim about the Fool's irrationality contrasts with Plato's Sophists, who claimed precisely that justice and morality are instrumentally rational, but only as a lesser good, i.e., only in those cases when the better options of coercion or deception are not available. Deception and coercion provide rational individuals with their highest profit. This is, I think, also the best way to understand the views on morality put forward by Alexander. He demands that our views on morality adjust to the fact that humans have been shaped by natural selection to further their own individual reproductive interests (1987, pp. 34ff.), amidst the conflicts that inevitably arise with the interests of other members of society. Cooperation with

others is only instrumental to self-benefits. This is evident, according to him, in the fact that everyone tries to shape cooperative interactions so as to profit more than their partners (pp. 102f, 109f.). In the same spirit, since moral systems are systems of indirect reciprocity and reputation is crucial, humans instrumentally cultivate a reputation as cooperative individuals (pp. 114, 191f.). Except, perhaps, for genuine psychological altruism between close kin, humans are psychological egoists focused only on their own benefit and cooperating only because coercion is not an option. In any case, they are always trying to influence and manipulate others so as to receive from them more than would be fair. This entails, of course, that humans will cheat if they can cheat undetected. In order for our deceitful natures to work effectively, natural selection has shaped us into self-deceivers as well (Alexander 1987, p. 123). We deceive ourselves into believing that we are good-natured, because if we did not, we would easily betray our deceitfulness and deceit would be ineffective. There is no genuine altruism; it is only a masquerade that we all play out so convincingly, that we have come to believe it ourselves. The philosophical idea that justice must be valued and is valued for its own sake (at least by some) is, according to Alexander, the product of self-deception.

Alexander's theory is an attempt to derive our deepest social-psychological nature from the ground-breaking theories on the evolution of social behaviour. These theories say that natural selection favours those genes that do their utmost to benefit their carriers. Alexander believes that this process has shaped our mind to direct all our behaviour, consciously or not, to the reproductive benefit of the agent. As stated above, he admits that humans can have non-instrumental desires for the welfare of close kin. But regarding non-kin, given the conflicts of interest that necessarily arise between them, all that humans need is the *appearance* of genuine altruism, although individuals may occasionally incur in large sacrifices for others, either by mistake or as victims of manipulation (pp. 104, 114, 191f.).

An interesting feature of his view is that he has a theory, inspired by Trivers (1971, 1985), for why our common-sense moral experience hides our basic selfishness. We deceive ourselves into believing that we are non-instrumentally interested in the welfare of others, so that we can better deceive them into the same belief (cf. the project for an evolutionary science of self-deception in Trivers 2011). The best way for unavoidable egoists to reap the benefits of cooperation is to build a reputation as fair players, because humans, both now and in our evolutionary past, value fairness over selfishness in social partners and prefer to interact with agents disposed to be fair. But if humans evolved to be motivationally selfish by biological design, there are no fair partners to choose from. Had this simple fact not been effectively concealed, large-scale human cooperation would never have evolved. The only chance for cooperation to evolve and prosper depended on the ability to conceal selfishness. Since conscious deceivers too often betray themselves involuntarily, deceiving others required deceiving oneself as well (Alexander 1987, p. 123). Thus, natural selection favoured the evolution of self-deception. A fundamental element in this self-concealment is that we denigrate selfishness and praise altruism in order to deceive ourselves and others into believing that we are, in fact, non-instrumentally (altruistically) interested in their welfare (p. 125). If we feel uneasy at a theory

that assigns this role to deceit and self-deceit in human morality, this feeling is just another symptom of self-deception. The theory thus explains the fact that altruism is high ranked in our system of values, while denying that humans are, or could be, genuine altruists. Alexander defuses in this way the testimony of everyday moral intuitions against the troubling claim that we are all egoists pursuing only our own benefit and willing to deceive others when it pays off.

If Alexander's view of our social nature is true, human cooperation rests on fragile foundations and the often quoted phrase authored by Michael Ghiselin 'Scratch an altruist and watch a hypocrite bleed' would exactly capture the frail nature of human morality. As Plato's Sophists claimed, humans cooperate only when the prospects of cheating and getting away with it are faint, while silently lurking for the opportunities where those prospects increase and a ruthless pursuit of individual advantage promises to pay. Obviously, this is no flattering picture of who we are. The world therein depicted is not one where most of us feel at home. Alexander would surely reply that we feel troubled by it because self-deception makes us think we are better. But I think he misses one reason why it should be troubling: he seems to think that cooperation is guaranteed as a stable expression of our deceptive and self-deceptive selves, but he is probably mistaken. In so far as it is only an expression of self-deceit, human cooperation is a façade that hides a manipulative agenda and a struggle for power. If his theory is true, no matter how real cooperation may look like, the struggle for power beneath it is the real master that determines our fate. If this is the world we live in, pessimists are fully entitled to have gloomy views about our future survival as a species. This is no argument against the theory itself, but it is one against the claim that the theory allows us to hope for a better future, or a future at all.

Fortunately, this apparently well-constructed theory has a fatal logical flaw, a contradiction in the way it conceives the evolutionary scenario. The starting point for evolution of self-deception consists in agents approaching others not only with selfish motivations, but also with a disapproval of selfishness that threatens to make cooperation impossible. If we disapprove of selfishness and know that everybody is selfish, how would cooperation even get started? The answer seems to be: only by hiding selfishness through deception and self-deception. Notice that agents disapprove of selfishness and approve of altruism *before the evolution of self-deception*. Self-deception is explained as having evolved under a selective pressure for genuine fairness in partners. But this character trait, given how natural selection works, cannot exist. So it seems that we valued genuine fairness before the evolution of self-deception. However, Alexander also makes self-deception responsible for the fact that we value fairness, because it is part of the strategy of concealing selfishness from consciousness. In sum, we value fairness as the product of self-deception, but at the same time self-deception evolved because we valued fairness and needed to cooperate for survival. But you cannot have it both ways.[2] The theory may work as an explanation of self-deception, but as such it requires the previous and

[2]This criticism of Alexander's theory was first argued in Rosas (2004).

independent existence of our commonsense values. Therefore, we cannot explain the latter as an output of the self-deceptive mechanism. We need an alternative explanation that does not undermine our everyday praise of moral agents as a form of self-deception.

2.5 Cooperation in a World of Selfish Agents

Evolutionary models explain why a trait exists by observing its fate in a population where the trait in question competes with rival traits. Simulating the dynamics of a population where agents of different strategies (given by competing traits) interact is a way of discovering the fate of the trait. A payoff matrix gives the utilities for all possible outcomes of the interactions between strategies and the strategies reproduce in direct proportion to obtained utilities. In this paper, I describe the evolutionary dynamics of social traits in a population as a thought experiment about cooperative or exploitative interactions, without the simulation tools. First we picture a population of selfish agents alone under the assumption of perfect mind-reading abilities. Then we relax this condition; and finally we introduce agents with social preferences in competition with agents that are selfish and lack those preferences. The result of the thought experiment will tell us whether morality and cooperation can subsist or not, supported only on selfish motivations and rationality (where natural selection replaces rationality in these models). If the extinction of cooperation follows in a population of rational egoists without social preferences, this provides evidence for the view that a motivation for treating others with fairness (a basic social preference) is a necessary requirement for morality.

The first two thought experiments are designed to illuminate the relative contributions of deceit and selfishness in a theory where morality is instrumental and eventually disappears. They show that human deceitfulness is far more important than selfishness in producing this outcome. In order to see this, imagine first a hypothetical world consisting purely of selfish agents, where deceit is impossible because intentions are transparent to everybody. In this case, intentions to defect will instantaneously be known by others. In interactions with a PD structure, perfect mind reading induces defection in all those that interact with latent defectors. Thus, if everybody is rational and is intent only on their benefit, it is better to form the intention to cooperate and carry it through. This brings the benefit of mutual reward, whereas second thoughts about not complying would be read by partners and would induce them to defect, resulting in the payoff for mutual defection. Since in a PD mutual defection is worse for both players than mutual cooperation, transparent rational egoists would always choose mutual cooperation. This is, in essence, Gauthier's argument to derive moral constraint from instrumental rationality (Gauthier 1986), only that he argues with the assumption of translucency instead of transparency. I shall show below that the argument does not work with translucency. But we can nevertheless acknowledge that cooperation would be the natural outcome in a world of egoists with transparent intentions, under the

assumption that everybody chooses what is best for them (everyone is rational). Selfishness can produce a perfect imitation of a moral world. Altruism would not exist, but neither would exploitation of others. Moreover, altruism would not be highly valued, nor would selfishness be disapproved of. Our common-sense system of values would not exist. The thought experiment shows us that our evaluative attitudes must result from the fact that we are not transparent, but translucent and sometimes opaque, and cannot always read correctly the intentions of others in cooperative interactions. Next, I show that in a world populated *only* by selfish agents who are not transparent, but rather translucent or opaque, cooperation is bound to disappear.

Translucency cannot support cooperation in a world of rational egoists. The crucial insight here is the link between translucency and the possibility of deceit. Imagine now that in the world of selfish, rational and transparent agents depicted in the previous paragraph, all agents change suddenly and become translucent. In contrast to transparency, translucency implies that agents are not infallible about others' intentions before action. There is a probability of misinterpreting their intentions or dispositions, although they have a better than random chance of correctly identifying them (this is Gauthier's definition of translucency). We can conceive this as a brute fact about the natural, involuntary signs of mental states through the body, without any deliberate conscious manipulation. Given this brute fact about the bodily expression of mental states, selfish agents can be expected to exploit translucency, 'engineering' misinterpretation in a specific direction: manipulate signs such that others believe to detect a cooperative intention where there is none. Translucency opens a door to the strategy of deceit, a development that Gauthier does not take properly into account (Sayre-McCord 1991). As agents gradually become better in deception, translucency is replaced with opacity or worse, because some agents succeed in putting a misleading appearance most of the time. Notice, however, that deceit does not make others believe in altruistic motives, for these do not exist in a world of selfish agents. Deceitful agents fake the intention to cooperate. This alters the probability of correctly identifying others' intentions, such that at some point it falls below random. At that point, it will no longer be true that cooperation with those you think will cooperate is the best move. Therefore, cooperation will likely disappear in a world of translucent agents that adopt deceit as a strategy. As soon as cooperation ceases, deceit loses its point. Deceit therefore, is the cause of the decline of cooperation. A disposition to denigrate both deceitfulness and selfishness could emerge along the process.

2.6 Fallible Mind Reading Makes Our Value System Emerge

In a world where agents are selfish and translucent, the evolutionary dynamics takes the population to a world where everybody uses deceit and fakes their character. In such a world full of deceit, the bare intention to cooperate is of no value. It is merely the strategic move of an agent trying to lure others into interaction only to

eventually exploit them. Those intentions would exist only as long as others can be deceived, as a transitory remnant of a world where agents had been transparent. To be able to explain the emergence of our praise of fairness, we have to change the initial composition of the population. The population cannot consist solely of selfish agents. Imagine instead that agents with a non-instrumental disposition for fairness are present in an important proportion. If these agents exist and are known to exist, *intentions* to cooperate can be valuable, not as such, but as expressions of the correct *motivations and character*. Since everybody judges a genuine disposition for fairness to be better than its absence, everybody tries to choose partners with that disposition. Consequently, deceitful agents try to fake being fair. Therefore, genuine as opposed to fake dispositions for fairness turn out to be valuable. In a world without transparency, where selfish agents can hide their deceitful plot, interacting with genuinely fair agents is the only guarantee of successful cooperation. And since cooperation is a reliable path to many goods, humans highly value a disposition for fairness. Selfish-deceitful agents survive as parasites of a system of values that praises fairness.

This evolutionary scenario requires the existence of fairness as a character trait simultaneously with the possibility of discriminating it among a population of selfish and deceitful agents. This is not a problem if selection favours a genetic linkage of the character disposition and its recognition in a population. Biologists picture natural selection favouring this linkage in the context of interaction between kin (Hamilton 1964), and then extending it to interaction with non-relatives. Axelrod and Hamilton (1981) speculated that kin selection could evolve into reciprocal altruism if altruism is conditionally expressed towards kin and kin identification focuses on cooperative behaviour. Given that cooperative behaviour can be used as a deceptive lure, recognition must go deeper than mere behaviour, i.e., behaviour must be taken only as *one* channel to character in a multichannel recognition system. If agents disposed to fairness are able to recognize their kind, then the benefits of cooperation fall only or mainly on agents that have a disposition for fairness. The recognition of fairness in others becomes a condition for the expression of cooperative behaviour. Those recognition abilities are a necessary condition for the evolution and preservation of cooperation. If these abilities are absent, genuine fairness is an easy prey to selfish characters. Their presence gives agents a critical adaptive advantage. A positive feedback loop emerges between both traits (fairness and its recognition in others).

Fairness exerts self-constraint: the agent avoids the use of deception as a means to force others into the role of exploited suckers. It dictates a conditional, but reliable cooperative attitude. The recognition of fairness of character is of crucial importance for co-operators and for selfish agents alike, because it is the only reliable avenue to the benefits of cooperation. The high value attached to this character reflects the difference in fitness resulting from interacting with agents having it, compared to interacting with agents lacking it. The adaptive value of detecting and choosing fair partners leads to a genetic predisposition to develop a preference for fairness. This is of course, just a hypothesis, but it is important to have one where the origin of our valuing altruism does not present it as self-deceptive. Both moral and selfish agents

value a moral character because it is a secure path to benefits. In contrast, selfishness becomes the target of aversive feelings because it is the vehicle for nasty, deceitful strategies of social interaction when mind reading is fallible. When deception is used in a social environment, choosing partners with a disposition for fairness is the best insurance against being deceived. It is therefore highly valued.

This model assumes that our mind-reading abilities, though imperfect, are accurate enough to support a strong psychological and social selection in favour of dispositions for fairness. This looks like translucency again, which we rejected above as an avenue for the evolution of cooperation. But recall that we rejected translucency in a world where all agents were selfish, where all have a strong reason to adopt deceit as a strategy as soon as any one adopts it. In this case, translucency can no longer guarantee that intentions will be correctly interpreted most of the time. The point here is that translucency cannot help the evolution of morality unless a substantial proportion of agents already have social preferences and thereby resist any temptation to manipulate translucency. Translucency cannot help if every agent adopts deception as their strategy, as rational egoists predictably would. Translucency will really help in the argument only when a substantial proportion of agents have a primitive disposition for fairness and will not adopt deceit as a strategy: their character revolts against it. They give honest signals of their character and thus create a chance for other agents to tap into objective differences between genuine and fake displays of dispositions for fairness. This raises the percentage of correct identifications. In this case, translucency is effectively equivalent to the idea that agents will have a greater than random chance of positively identifying moral or selfish dispositions when they are really there. The natural consequence is that a social-psychological selection for genuine fairness of character can take place in human evolution. In this respect, this informal model preserves one of the insights that make Trivers's defence of reciprocal altruism relevant for morality: in cooperative enterprises partners are judged and chosen in virtue of their motivational dispositions, which are the only reliable signs of a consistently cooperative attitude (Trivers 1971, pp. 50–51; see also Rosas 2007; Nesse 2007). In this way, the benefits of cooperation are circumscribed to those that are non-instrumentally interested in the welfare of others. I have here developed Trivers's 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 genuine dispositions for fairness.

References

Alexander, R. 1987. *The biology of moral systems*. New York: Aldine de Gruyter.

Axelrod, R., and W.D. Hamilton. 1981. The evolution of cooperation. *Science* 211: 1390–1396.

Bowles, S., and H. Gintis. 2011. *A cooperative species; Human reciprocity and its evolution*. Princeton/Oxford: Princeton University Press.

Buss, L. 1987. *The evolution of individuality*. Princeton: Princeton University Press.
Darwin, C. 1981 [1871]. *The descent of man and selection in relation to sex*. Princeton: Princeton University Press.
Fletcher, J.A., and M. Doebeli. 2009. A simple and general explanation for the evolution of altruism. *Proceedings of the Royal Society B: Biological Sciences* 276: 13–19.
Frank, R. 1988. *Passions within reason*. New York: W.W. Norton.
Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.
Gibson, M. 1977. Rationality. *Philosophy & Public Affairs* 6: 193–225.
Gintis, H., S. Bowles, R. Boyd, and E. Fehr. 2003. Explaining altruistic behavior in humans. *Evolution and Human Behavior* 24: 153–172.
Hamilton, W.D. 1964. The genetical evolution of social behavior I and II. *Journal of Theoretical Biology* 7(1964): 1–52.
Hampton, J. 1991. Two faces of contractarian thought. In *Contractarianism and rational choice*, ed. P. Vallentyne, 31–55. New York: Cambridge University Press.
Hobbes, T. 1651. *Leviathan*, ed. C.B. Macpherson. London: Penguin.
Hume, D. 1777, 1902. An enquiry concerning the principles of morals. In *Enquiries concerning the human understanding and concerning the principles of morals*, ed. L. Selby-Bigge, 169–285. Oxford: Oxford University Press.
Joyce, R. 2006. *The evolution of morality*. Cambridge, MA/London: MIT Press.
Lehmann, L., and L. Keller. 2006. The evolution of cooperation and altruism – A general framework and a classification of models. *Journal of Evolutionary Biology* 19: 1365–1376.
Maynard Smith, J., and E. Szathmáry. 1997. *The major transitions in evolution*. New York: Oxford University Press.
Michod, R.E. 2007. Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences* 104(S1): 8613–8618.
Nesse, R.M. 2007. Runaway social selection for displays of partner value and altruism. *Biological Theory* 2(2): 143–155.
Rosas, A. 2004. Mindreading, deception & the evolution of Kantian moral agents. *Journal for the Theory of Social Behaviour* 34(2): 127–139.
Rosas, A. 2007. Beyond the sociobiological dilemma. Social emotions and the evolution of morality. *Zygon* 42(3): 685–699.
Rosas, A. 2010. Beyond inclusive fitness? On a simple and general explanation for the evolution of altruism. *Philosophy and Theory in Biology* 2: e104.
Rosas, A. 2012. Disentangling social preferences from group selection. *Biological Theory*. doi:10.1007/s13752-012-0013-y.
Sachs, J.L., U.G. Mueller, T.P. Wilcox, and J.J. Bull. 2004. The evolution of cooperation. *The Quarterly Review of Biology* 79: 135–160.
Sayre-McCord, G. 1991. Deception and reasons to be moral. In *Contractarianism and rational choice*, ed. P. Vallentyne, 181–195. New York: Cambridge University Press.
Sober, E., and D.S. Wilson. 1999. *Unto others: The evolution and psychology of unselfish behavior*. Harvard: Harvard University Press.
Tinbergen, N. 1963. On the aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20: 410–433.
Trivers, R. 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology* 46: 35–57.
Trivers, R. 1985. *Social evolution*. Menlo Park: The Benjamin/Cummings Publishing Company.
Trivers, R. 2011. *The folly of fools. The logic of deceit and self-deception in human life*. New York: Basic Books.
West, S.A., A.S. Griffin, and A. Gardner. 2007. Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology* 20: 415–432.