ISSN: 2282-4189

LUISS

Research Center
for European Analysis
and Policy

EMUNA Brief 1/2026

SCIENCE AND ETHICS IN DIFFERENT CULTURES – 9 February 2026

Giorgio Vallortigara

A Biological Origin for the Moral Sense?

A Biological Origin for the Moral Sense?

Giorgio Vallortigara•

Abstract

After briefly outlining some of the evidence suggesting the presence of a proto-morality and a sense of justice in infants of our species and in adults of nonhuman primates, this paper considers the theoretical foundations underlying the biological evolution of cooperation and altruism. In particular, it is emphasized how the mechanisms of reciprocal altruism have led to (or required) a sophisticated psychology that includes the analysis of the costs and benefits of actions, along with the ability to recognize individuals and retain their characteristics in memory over even prolonged periods.

• Centre for Mind/Brain Sciences, University of Trento

I invite you to consider the following phrase and try to guess to whom it belongs: "In the transmission of values and in teaching the difference between good and evil, the teacher can never replace the priest or the pastor." I imagine you might think of the name of some Pope or equivalent figure. Instead, you will be greatly surprised to discover that the phrase was made public in 2007 by Nicolas Sarkozy, a former president of a state with a highly pronounced form of secularism. It is a very widespread point of view, according to which moral sense derives from some form of social and cultural indoctrination linked to religion. It is found in famous phrases such as B. Franklin's "Religion will be a powerful regulator of our actions... and render us benevolent, useful and beneficial to others or in F. Dostoevsky's apocryphal "If God does not exist, everything is permitted."

Regardless of one's beliefs, there is an interesting empirical question here for those studying behavioral biology: whether moral principles can actually be the result of some kind of social, religious, or other indoctrination, or whether they are already hardwired into the brains of biological organisms, albeit in a rudimentary form. Today, we have evidence supporting the second hypothesis on two fronts. One is research with very young children. Karen Wynn and Paul Bloom from Yale, using a classic habituation/dishabituation technique, showed children videos of a circle struggling to climb a mountain, followed by a good triangle that gets underneath and helps push it, or a bad square that pushes it down from above with violent blows. At the end, the children are given the opportunity to choose one of the two characters from a tray, and they choose the triangle. Infants a few months old have a natural, spontaneous preference for helpful, good, and cooperative objects. Then there are studies on nonhuman primates. One of the most famous, conducted by Franz de Waal, demonstrates the rudiments of a sense of justice in monkeys. It shows two monkeys trained to give a token to the experimenter in exchange for a reward. The reward can be something they like a lot, like a grape cider, or something they like less, like a slice of cucumber. Each animal can observe what happens to its partner who performs the same actions. One of the animals notices that while its partner is rewarded with the grape for the correct action, it is rewarded with the cucumber slice. The animal's reaction to this injustice is quite lively, even throwing the cucumber slice at the experimenter to reject him.

Perhaps, therefore, in transmitting values and learning the difference between good and evil, both the teacher, the priest, and the pastor rely on a natural predisposition toward moral behavior that we share with other creatures, and which is therefore very old, and which appears very early in infants, thus in the absence of specific indoctrination. The next question we must ask is where all this comes from, how organisms develop this predisposition toward moral behavior.

Karl Popper stated once that scientists should be scholars of problems, not disciplines. Indeed, there are problems that cut across disciplines. One such problem is the origins of cooperation and altruism, which has been and is a source of concern for scholars of political science, law, economics, and human ecology, but also for biologists. Since Darwin's time, biologists have been faced with the problem of understanding how it is possible that entities whose behavior is aimed solely at maximizing their own benefit—their benefit, in this case, is measured by the copies of their genes in the next generation—can develop altruistic and cooperative behaviors. As is known, animal behavior can result in horrific atrocities. An example is the behavior of the female praying mantis, which, after mating, proceeds to cannibalistically feed on her partner. This may seem morally objectionable, but it is completely understandable from the perspective of natural selection: the female needs energy resources and has

them readily available in the male's body, so she takes them—something very similar to Keynes's famous free lunch. The most surprising part of the story is that the cannibalistic meal often begins before copulation. There is a reason for this, too, a neurobiological one this time. The cranial ganglion of insects—like the cortex of vertebrates—is essentially the seat of inhibitory mechanisms, which means that when you remove it, the stereotyped basic motor functions are maintained, even made even more efficient. In effect, what happens is that the headless male praying mantis becomes an excellent copulator. On a humorous note, women often say that even in the human species brainless males function better. Examples of selfish behavior in the animal world are countless. At the same time, Darwin was aware that alongside these selfish brutalities, altruistic and cooperative behaviors are often observed. Consider soldier bees sacrificing themselves at the entrance to the hive for the other individuals in the group. The bee that stings you is soon doomed to die because the hook-shaped stinger, when extracted, tears out the animal's abdominal organs. The same is true for passerines, which send out an alarm call and thus save their companions, but by attracting the predator's attention, they put their own lives at risk. So, what is the problem? If we assume that these behaviors have a genetic basis—that is, if we imagine, to put it crudely, that there is a gene for sacrificing oneself for others—then by definition, the individual carrying this gene will not leave copies of that gene, and therefore the gene should disappear. This is somewhat of the problem that Darwin had already intuited and that dominated behavioral biology studies until, in the 1970s, William Hamilton proposed an initial solution, the so-called kin selection. Hamilton's idea is simply that when we consider the hard currency of natural selection—the number of gene copies left in the subsequent generation—the calculation should be made not only with reference to the individuals who perform the altruistic action, but in general to all the gene copies contained in those bodies and in the bodies of other individuals to whom they are more or less closely related. There is a famous story told of the biologist Haldane, to whom a journalist once posed a question: "Professor, would you be willing to throw yourself into the river to save someone from drowning?" To which Haldane replied, "I would be willing to do so if there were at least three children in the river, or five aunts, or at least nine first cousins..." What did Haldane mean? He was saying that since in diploid species, like ours, we share 50% of our genes with our offspring, if I jump into the river to save two children, I'm 50-50 even, but if I save at least three, it's worth the risk. Haldane and Hamilton are obviously not arguing that people do genetic calculations. They are saying instead that there may be mechanisms, natural predispositions, that if they make you behave that way, the purely mechanical result is that there will actually be more copies of that gene for that behavior in the next generation. It all happens without anyone wanting it and without anyone knowing genetics. The mechanism is the blind mechanism of natural selection.

Hamilton's theory of kin selection has been crowned with enormous successes, particularly in the study of organisms characterized by a special sociality, such as social insects, bees, for example. Now we can try to develop—and in Italy it would work particularly well—a theory of universal nepotism and view altruistic acts in terms of kinship. However, biological organisms often display altruistic behavior even toward individuals to whom they are not genetically related. And the question is: why do they do it?

An intuitive explanation, part of the popular wisdom of certain old nature documentaries, is that they do it for the good of the group, of the species. This is a viewpoint that has dominated classical

ethology; indeed, you can find many references to the good of the species in the writings of Lorenz, for example. However, this hypothesis is flawed from the perspective of the standard theory of natural selection. To illustrate my point, let me give you a classic example. There is an old story about the behavior of Arctic lemmings. Lemmings are rodents that, since they are subject to population explosions, can find themselves with insufficient resources for the entire population. It is said that some of them sacrifice themselves by throwing themselves off cliffs and drowning in the sea: they sacrifice themselves to save the group, for the good of the species. This tall tale was invented by Disney documentary makers when they made a famous documentary, later debunked, in which the poor lemmings fall off cliffs.

Intuitively, we all have the idea of group selection in which an individual can sacrifice himself for the good of others in the species. We need to understand why it does not work. It does not work from a mathematical standpoint, and you can guess why by looking at this classic Gary Carson illustration.

You can see the lemmings jumping into the water, but there is one traitor who decides to don a lifebuoy. If you have a hypothetical gene for jumping into the water, it could happen that, through

mutation or from another population, its rival allele arrives and says: "I'm not going to jump into the water!" Obviously, the individual carrying this gene will have enormous reproductive success because it survives and leaves behind copies of its genes, including the gene that says not to jump into the water. In a short time, a hypothetical population of altruists will therefore be invaded by selfish genes. This is why the group selection theory does not work, even though it is very, shall we say, politically correct. One way to conceptualize these matters is the famous prisoner's dilemma. The following a simplified version that goes straight to the point, to the relation with altruism.

Imagine you are a jewel thief. You have a suitcase full of jewelry and need to turn it into hard cash. You arrange with a fence to meet in Rome at midnight in Piazza Navona. You need to go with the suitcase full of jewelry and the fence with the suitcase full of cash. Now the question is: what's better for you? Go with an empty suitcase or a full suitcase? Think about it for a moment and it is clear that if you go with a full suitcase and the other person arrives with a full suitcase, you have both made a good deal. But can you trust him? If you go with a full suitcase and the fence arrives with an empty suitcase, you have lost everything. So the best thing to do is go with an empty suitcase. If he has a full suitcase, you get the famous free meal; if he also has an empty suitcase, well, you both get nothing. It is a shame, because it would have been better to cooperate, but the problem is how to trust him. There is the problem of free riders, of cheaters, of the lemming who decides to don a lifebuoy. In this version there is no solution other than defection: the best thing you can do, the best thing a rational creature can do, is not to cooperate. But this is a very unrealistic situation in which everything ends in a single meeting. A more realistic situation in the same game is the following. You are a professional thief, and so every month you decide to meet the fence in Piazza Navona, and every month you make the exchange. So at this point, in the so-called iterated version of the game, the situation changes a bit because one says: Well, what do I do the first time? It is not nice to go there for the first meeting and start with an act of distrust, that is, to show up with an empty suitcase. Maybe it is better to go there with a full suitcase and then see what happens. Now, this iterated version is a difficult problem, with no mathematical solution, but it was addressed with an experimental mathematical technique. Years ago a tournament was organized in the form of a game. It was a round-robin tournament in which computer programs competed and could cooperate or defect using the form they preferred. Programs developed by highly sophisticated expert mathematicians competed, but the surprising thing is that the winning program, formalized by Anatol Rapoport, is very, very simple, just two lines of Basic at the time, Tit for Tat, based on the idea that you always start by cooperating and then, like a mirror, you mirror what your opponent does. So, if your opponent cooperates, you continue to cooperate; if they defect, you immediately punish them with a defection. Once Tit for Tat establishes itself as the majority strategy in a population, it becomes what is technically called an evolutionarily stable strategy, that is, a strategy that cannot be displaced by rival strategies. Tit for Tat is taught today by diplomacy experts as a basis for negotiation: when negotiating, do not start the deal with a hard face; try to be friendly and cooperative. However, if someone deceives you, do not wait, do not dwell on it, saying, "Let's see if they do it again." No, immediately punish the person who did not cooperate. On the other hand, do not sulk; if the other person cooperates again, you too should cooperate. These are very simple principles, but they also form the basis of the theory developed by Robert Trivers, the theory of Reciprocal Altruism, which describes how biological organisms can develop forms of altruism and cooperation in the absence of genetic

kinship. This is the famous principle of "I scratch your back, tomorrow you scratch mine." This idea of reciprocal altruism is interesting because it has implications for the sophistication of the psychology of those who adopt it, that is, the cognitive processes required to implement it. To implement the strategy of reciprocal altruism, one must be able to evaluate the costs and benefits of the act, but this was obviously also true for kin selection. Furthermore, however, it requires a good memory and the ability to recognize individuals, meaning that you must remember that it was Paolo, not Giuseppina, the individual towards whom you performed the altruistic act, and therefore it is from Paolo that you expect reciprocity, that is, that he return the favor. And obviously you have to have a good memory because you have to remember Paolo and Giuseppina and remember when all this happened. The implications for morality are also quite obvious, because one way to combat free riders and cheaters is to develop brutal methods of punishing anyone who breaks the rules. A more subtle, more sophisticated approach is to develop moral rules for this purpose, rules that perhaps have to do with what is called reputation. Much of what humans do is tied precisely to the fear of losing our reputation, which is a way to maintain this kind of reciprocity that characterizes us as human beings.

What distinguishes us from other primates is precisely that even trivial forms of cooperation are very difficult for them to achieve. An example is the ultimatum game. There is a certain amount of money, say 100 euros, and the game consists of proposing an equal division of the sum to the other player. The other player can say "I accept" or "I refuse." If they accept, we each get our share; if they refuse, we both lose everything. So, for example, I'm playing with one of you and I say, "Look, we have €100. How about I get €70 and you €30?" Usually, the other player accepts. But if I behave very unfairly, and I give myself €99.99 and the rest to you, the other player says no, deems me to be behaving badly, and therefore punishes me. This is completely unreasonable in terms of classical economic theory because even a cent should be accepted since it is better than nothing. But humans do not do this because they possess what we call a sense of justice. Chimpanzees, on the other hand, behave rationally in the ultimatum game, accepting even the smallest prize.

An interesting question is whether we are the most altruistic animals on Earth. The answer is no. There are animals capable of pure, selfless altruism, even greater than ours. For example, a truly astonishing behavior has been documented in African grey parrots. In one experiment the parrot was trained to give a token to the experimenter in exchange for a reward. During the experiment, there were two parrots: one had access to the tokens but no way to provide them to the experimenter; the other could interact with the experimenter but did not have access to the tokens. What happened was astonishing, something you would never observe among monkeys and, I believe, would be difficult to observe among humans. The animal with access to the tokens provides them to its partner, who then receives the reward. It is interesting to ask how this type of behavior evolved in this species. The reason, apparently, is that African grey parrots maintain a very close monogamous relationship. So, when you have a lifelong partner, it does not matter if they can occasionally achieve more than you with less effort. Because in any case it will be the offspring of both of you who will benefit. Interestingly, the rule, almost automatic, does not require the other to be a partner; it can even be a stranger of the same species. This does not apply, as we well know, to species like ours.