Some controversies around Moral Nativism

Authors: Roger Rex (University of Brasilia); Paulo Abrantes (University of Brasilia)

Abstract
The theory of evolution sparked a series of questions about the origins of moral judgments and the underpinning principles. In particular, it reinforced the debate about moral nativism. In this paper we scrutinize two research programs that advocate respectively the existence of an innate ability to judge morally and a predisposition to moralize behaviors with certain contents. The best-known version of moral nativism argues for the existence of a moral grammar, by analogy with the Chomskyan model of principles and parameters in linguistics (Universal Moral Grammar). The second program argues for the existence of a moral domain, i.e., of a small set of moral intuitions found in all societies (Moral Domain Theory). We critically evaluate the arguments commonly used to ground both theories: ease of learning even in face of poverty of stimulus; the principle of double effect; the pre-established order of moral development in individuals; universality and antiquity of the moral phenomenon. We are concerned with how they fare at relevant contemporary research in Cognitive Science and meet constructivist arguments proposed by Jesse Prinz and Kim Sterelny, among others. We found out that there is little evidence that our moral judgments follow the model of principles and parameters, although it can be useful as a heuristic device for guiding future research. At the same time, ease of learning suggests that the human brain is somehow prepared to learn moral rules and that the types of rule we adopt are constrained by our biology. Furthermore, the fact that the capacity to make moral judgments develops according to a similar schedule in different cultures indicates that it is an endogenous one. Although it depends, for sure, on the existence of culture, it is not reducible to cultural phenomena. Based on a distinction made by Darwin between social instincts and intellectual powers we argue that thinking in a normative way requires being aware of the grounds of our beliefs and actions, and this implies self-consciousness. It is difficult to discern if morality is an adaptation or just a spandrel. Even if it might have emerged as a side effect in our evolutionary history, it might have
undergone subsequent adaptive structural changes, possibly because of interactions with cultural dynamics.

1 Universal Moral Grammar

The most discussed and well-known version of moral nativism employs language as a model to explain our moral cognition, conveying the idea that there is a universal moral grammar (Dwyer, 2006; Hauser, 2006; Mikhail 2011; Hauser et al., 2008). According to this theory the morality of each individual is the result of a combination between a universal set of principles and parameter settings emerging from cultural interactions (Hauser et al., 2008, p. 122). The main arguments used to support the existence of a universal moral grammar are focused on the poverty of the stimulus. According to this argument, there is a problem in explaining a cognitive ability when there is a gap between the complexity of the learning object and the resources available to the learner. The existence of this gap leads to posit some kind of innate structure responsible for the learning ability (Sripada, 2008, p. 325-6).

Defenders of the universal moral grammar (UMG) theory argue that the behavior and the moral discourse of adults, to which children are exposed, would not include sufficient information to allow the acquisition of a moral competence unless there are learning mechanisms dedicated exclusively to the moral domain (Prinz, 2008, p. 430; Joyce, 2006, p 137). The first difficulty with this argument is the fact that moral principles, as a rule, are made explicit. This can be compared to language acquisition: in this domain children do not receive a significant amount of explicit negative information because different languages are full of enigmatic principles which the language speakers don't have conscious access to. In general, ungrammatical sequences are not produced and labeled as deviant, they are just not produced. With respect to morality, forbidden acts are produced and described, and the prohibitions are made explicit by means of intensive instruction (Sterelny, 2010, p. 290-1).

However, Hauser and Mikhail believe that moral judgments are also mediated by principles inaccessible to our consciousness. They argue for this thesis based on the research with the trolley problems (Hauser et al, 2008; Mikhail, 2011, p. 38). Hauser and Mikhail think that the respondents apply the principle of double effect to find

---

1 This is how Gilbert Harman describes the principle of double effect: it is worse to cause harm to someone (who has not consented to this) as (part of) your means to bringing about a greater good to
answers to these dilemmas, although they don't know how to formulate the principle. Since this principle is rarely articulated and it is not an obvious generalization from particular cases, it is hard to conceive how it could be learned. If people do not know it consciously, they are not able to teach children about it. If children nevertheless learn it, they seem to be somehow prepared for this (Roedder & Harman, 2010, p 286-7). The alleged solution is that principles such as these are inscribed on the morality of each individual as part of a universal moral grammar (Harman, 2000, p. 225).

Constructivists, on the contrary, do not interpret the decisions taken by the respondents as the application of unconscious rules or principles specific to morality. Prinz, e.g., believes people learn that killing is wrong through paradigmatic cases and they are more tolerant of behaviors that deviate from those (Prinz, 2013, p. 106). The amazing convergence of answers to the classic *trolley problem* could be explained by the fact that the solution considered morally forbidden - pushing the fat person onto the tracks - combines a set of characteristics normally considered reprehensible in different societies: the fat person is being employed as a means to an end, the case is much like the paradigmatic cases of murder and there is a strong personal element. The act of pushing someone is much more personal than the act of pulling a lever, which encourages more people to consider it impermissible. When these factors are isolated, the participants fail to converge towards a common response. This fact becomes salient when we analyze the results of research involving two scenarios called "Loop Track" and "Man-in-Front" by Mikhail (2011, p. 107-8). The essential difference between the two scenarios is that in the first one the death of the person located on the adjacent rail is a necessary means to prevent the deaths of five others; while in the second case, the death of that person is just a side-effect of diverting the train. In the first scenario, 48% of respondents considered diverting the train permissible; in the second situation, 62% felt this way. Although the answers diverged greatly, Mikhail believes that the principle of double effect was applied since the majority of participants judged impermissible the act that caused injury as a means for a greater good.

---

2 Regarding the classic version of the dilemma, originally formulated by Philippa Foot, the vast majority of respondents replied it was morally permissible to divert the trolley to an adjacent rail to save the lives of five workers, even if that conduct caused the death of an innocent. However, they did not consider permissible pushing a person in front of the trolley in order to save those same five workers. Just a few respondents could explain why only one of these behaviors was permissible.

3 For the distinction between personal and impersonal acts, see Greene (2001).
In spite of this, the question that seems important here is the extent to which we can, from such divergent results identify universal principles that purportedly have produced the moral judgments observed. The fact that it is possible to come up with principles such as the double effect to summarize the moral intuitions of a large part of the respondents does not mean that their intuitions have been produced by means of the alleged principles. Therefore, it is important to distinguish between an external and an internal approach. The external approach seeks to develop a set of principles which explain most of the intuitions expressed by the participants, and it is always possible to elaborate more than one set of such principles that can fulfill the explanatory task equally well. The internal approach, on the other hand, seeks to identify the principles that are causally responsible for producing the observed intuitions. Nichols points out that UMG supporters use the trolley problems to come up with a set of principles that is consistent with most intuitions without bothering to show that those principles are actually involved in the causal production of the intuitions (Nichols, 2005, p. 360).

The trouble, that the intuitions of many people are not compatible with the supposed universal principles, is circumvented by appealing to the distinction between competence and performance\textsuperscript{4}. Mikhail and Hauser believe that certain psychological limitations, performance errors, emotions and other factors distort the moral judgments. According to them, the solution is to isolate those factors and to identify the moral judgments that exhibit the optimal operation of the moral faculty (Mikhail, 2011, p. 103). However, it is unclear whether what they consider interference, especially certain emotions, is not in fact constitutive of our morality.

There is a difficulty in combining the idea that moral judgments conform to a grammatical structure and the idea that emotions have a causal role in the production of intuitions. Mikhail, Hauser and Dwyer, while recognizing the existence of an important interface between cognition and emotion, choose to argue that moral judgments cause emotions and not vice versa. Confronted with empirical evidence that we can modify moral judgments through the manipulation of emotions by means of hypnosis and environmental changes (Wheatley & Haidt, 2005; Schnall, Haidt, Clore & Jordan, 2008), they seek to preserve their theories appealing, again, to the distinction between competence and performance: they argue that emotions just affect moral performance, but are not constitutive of the moral faculty itself. Emotional processes sparked before

\textsuperscript{4} For the competence and performance distinction, see Chomsky (1965, p. 3-4)
the computational processes required to produce moral judgments can affect the latter and motivate certain actions, but are not necessary for the functioning of the moral faculty (Dwyer, Huebner & Hauser, 2010, p 495). There is, however, evidence that emotional responses are required for a normal moral development. The lack of empathy in children with psychopathic tendencies, e.g., compromises this development (Blair et al., 2006, 1997; Dupoux & Jacob, 2007, p 376).

Yet in order to ward off the idea that we have conscious access to the reasons underlying our moral judgments, some advocates of UMG argue that moral justifications are not connected with moral judgments. Dwyer (2009), e.g., points to the phenomenon described by Haidt (2001) as moral dumbfounding: under certain circumstances – e.g., dilemmas about incest - people simply cannot find reasons to justify their moral judgments. Defenders of UMG believe that this kind of phenomenon occurs because the principles that led to the decision of these people (operative principles) are inaccessible to consciousness (Mikhail, 2011, p 83; Dupoux & Jacob, 2007, p 374). According to Haidt we overestimate the importance of conscious reflection on our moral judgments. In most cases this conscious process serves only to rationalize intuitive judgments. However, Haidt points out that reflection is not just an epiphenomenon of morality. In some circumstances it can modify over time the intuitive judgment of the individual, and in many cases it serves to change the intuitions of other group members (Haidt & Bjorklund, 2008). The disgust vegetarians have for meat, e.g., is a consequence, not a cause, of their moral convictions (Fessler et al., 2003). Therefore there might be, in the context of morality, a diachronic interaction between intuitive reactions, on the one side, and conscious and articulated thought, on the other side. Explaining the interaction between these two kinds of cognition is a major challenge faced by UMG theorists, and they do not have a convincing model to propose. The analogy between language and morality cannot help here, because the processing of syntactic rules does not rely in any way on conscious reasoning (Sterelny, 2010, p. 282-4).

Hauser, Dwyer and Huebner, even when they recognize that children receive explicit instructions about morality, argue that they are not sufficient to explain the sophisticated moral rules found in the judgments children make. In particular, the imperatives they hear - "always keep your promises" - would not be enough to explain how children are able to identify exceptions to these rules (Dwyer, Huebner & Hauser, 2010, p. 492). Still, the explicit instructions are not as rough as Hauser, Dwyer and
Huebner believe. Parents usually offer explanations for their instructions. They resort to different reasons and norms to justify why the child should change her behavior. In many occasions, parents seek to induce her to adopt the perspective of the offended person, indicating the suffering caused and making it clear that the child was responsible for that harm (Hoffman, 2001, p. 143).

In addition to explicit instructions, children receive other stimuli. Stories, myths and children's songs are full of moral lessons (Sterelny, 2010, p. 288-9). Moreover, children’s lives are not passive. They do not just listen to stories and watch what others do. They act and face conflicts with other children involving harm and distribution of resources. Perhaps children would not be able to learn the rules for distribution of goods just by observing the behavior of their close relatives; but they often take part in discussions in which they seek to find a suitable rule for resolving this type of conflict - e.g., *I divide and you choose*. Interactions such as these are part of the learning process and are loaded with information related to morality (Sterelny, 2010, p. 288-91).

In short, it is not difficult to show that explicit and implicit moral stimuli are plentiful in the environment in which children develop. However, this is not sufficient to refute the poverty of the stimulus argument. The refutation requires a demonstration that the information received is rich enough to explain the capacity developed, since the poverty of the stimulus is not only related to the amount of stimuli, but also to the complexity of what is learned.

Concerning morality, the subject to be learnt seems to be much simpler than in the case of language. Moral norms are not as obscure as the recursive and structural rules of grammar. The development of a moral capacity requires learning a set of more specific rules, e.g., *share your toys, do not hit other children, respect your elders, etc.* (Sripada, 2008a, p. 328). Notwithstanding, defenders of UMG employ another argument to support the complexity of the task of making moral judgments: how can we explain the development of a capacity to judge an infinite number of cases from a finite experience?

Scholars that seek to explain morality would also face, like the linguists, the *projection problem*, i.e., the problem of explaining how individuals are able to apply their moral knowledge to cases different from those they experienced previously (Mikhail, 2011, p. 30). A mature speaker of any language has been in touch with only a limited number of sentences throughout her life. Despite this, she is able to build an infinite number of sentences that are comprehensible to other speakers. Given the
limited storage capacity of our brains, it is not possible that all these sentences are stored individually. So there must be a grammar by which we can, from a limited vocabulary and sentence patterns, build an unlimited number of expressions (Mikhail, 2011, p 45-6; Dwyer, Huebner & Hauser, 2010, p 488-9).

UMG enthusiasts propose that something similar takes place in morality: an individual with a developed sense of justice would be able to make an unlimited number of intuitive judgments about the moral properties of actions and agents (Dwyer, Huebner & Hauser, 2010, p. 489). Again, considering the limited storage capacity of the brain and the unlimited ability to make judgments, it follows that the moral faculty is more than a simple list. There would be a cognitive system composed of principles and rules responsible for producing an unlimited number of moral judgments (Mikhail, 2011, p. 46 and p. 72).

There is, however, a big difference between morality and language: to explain how we are able to interpret completely new phrases from the sounds we hear, we have to appeal to a complex set of principles, since there is a difficulty in explaining the reversibility in language. How is it possible that someone can produce a new set of sounds from semantic representations and someone else can rebuild these semantic representations based on that combination of sounds never heard before? When a native speaker makes a judgment about the grammaticality of a sentence, she does not judge only if that sentence is correct or incorrect, she also seeks to understand (rebuild) what the other person meant. In moral judgments there is nothing similar to this reversibility. While language is the result of a generative system, morality is simply the outcome of an evaluative system (Dupoux & Jacob, 2007, p 376). This means that morality looks more like our ability to evaluate the taste of a food, or the quality of a work of art, than our capacity for language.

In this context, constructivist theories based on the ability of pattern recognition have the advantage of simplicity over UMG, as they seek to explain moral learning through mechanisms employed in other forms of knowledge. The first step to postulate the existence of a specific moral learning capacity would be to demonstrate the impossibility of learning to make moral judgments through general-purpose learning systems, such as pattern recognition. No one doubts we have this ability. Thus, before proposing the existence of a special-purpose system, we should explain why general-purpose learning systems are not able to provide an adequate explanation (Prinz, 2012, p. 128 and 150).
The human mind is good at recognizing patterns and noticing similarities between different situations. The exercise of this capacity results in intuitive judgments about new cases. Pattern recognition can be extended to infinite new cases and occurs rapidly and automatically. In addition, the person who recognizes the pattern often cannot explain which features motivated what she recognized. An expert on birds, for example, can identify at a glance certain species, even when she is not able to explain exactly how she did this (Sterelny, 2010, p. 287-8). So, if moral intuitions were the result of a pattern recognition process, the fact they are fast and that we have difficulty in distinguishing their causes, would not be a surprise (Sterelny, 2010, p. 288-9).

Sterelny sees the relationship between tacit and explicit principles present in morality as something very similar to what happens with other skills acquired through a general-purpose learning system. An artisan, for example, has considerable explicit knowledge - she can easily explain many of the techniques she uses - which coexists with implicit knowledge resulting from the habit of practicing her profession and with a capacity to recognize patterns intuitively (Sterelny, 2012, p. 132-42). However, this distinction between explicit knowledge and tacit knowledge, both in the moral context and in relation to artisan activity, is not very sharp: the artisan may be able to explain why she chose or rejected a certain material, but this reconstruction ends up being partial because she does not have access to all the reasons that really motivated the choice, which means that explicit and implicit knowledge are interwoven (Sterelny, 2010, p. 293-4).

If moral judgments can be explained by pattern recognition processes, the main arguments of UMG proponents collapse, including the poverty of the stimulus, and also their solution for the projection problem. They must show that the subtleties and abstractions involved in moral judgments make that explanation implausible. So far, efforts employed by Hauser and Mikhail to exemplify the subtleties involved in moral judgments have focused on issues involved in our abilities to orient ourselves in the social environment, such as our ability to attribute mental states and intentions to others and to mentally represent the characteristics of an action. But these abilities are not specific to morality (Sterelny, 2012; Sterelny, 2010, p 287-8). Their general character makes a case against UMG’s claim that we have a specific moral faculty.

2 Moral Foundations and Affective Resonance
Morality exists in all human societies we know of and almost every individual develops a sense of it without formal instruction and without intentional effort (Joyce, 2006; Ayala, 2010, p. 2). Haidt elaborated his Moral Foundations Theory (MFT) based on these universal features. Together with Joseph (2004), he sought to identify common principles of human morality through an analysis of five studies on universal characteristics (Brown, 1991; Fiske, 1992; Schwartz & Bilsky, 1990; Shweder et al, 1997; de Waal, 1996). After trying to list all things humans and chimpanzees seemed to value in the behavior of others, they summed up their findings in five categories: (a) sensitivity or aversion to pain signals and suffering in others (harm/care); (b) negative responses to those who fail to reciprocate favors (fairness/reciprocity); (c) anger against those who fail to show signs of deference and respect (authority/respect); (d) emotions related to disgust, required to explain moral rules about food, sex, menstruation and disposal of corpses (purity/sanctity); (e) attitudes towards group boundaries (in-group/out-group).5 Haidt claims that these five groups of intuitions are moral foundations (Haidt & Bjorklund, 2008, p 202-3; Haidt & Kesebir, 2010, p 822). He argues that each one represents a mental module and is connected to different families of emotions: Suffering leads to empathy and compassion; disregard for hierarchy produces resentment and contempt; violations of reciprocity generate anger and guilt, violations related to purity provoke disgust. These domains are supposed to be universal, but each culture can define their specific contents (Haidt & Joseph, 2004).

Haidt wondered if our capacity to make moral judgments relied on something similar to taste receptors. Cuisines are cultural products and each one is unique and has a set of key ingredients. Even so, they are constructed on the basis of a sensory system which includes only five types of taste receptors. The structures of our tongues, noses and brains restrict the number of possible types of cuisines, but leave plenty of space to creativity (Haidt & Bjorklund, 2008, p 201-2). Haidt’s thesis is that there is something similar in morality. That is, although the moral matrix of a given group is a cultural construct influenced by various particular historical and environmental factors, it must be compatible with minds equipped with certain types of social receptors (Haidt, 2012, ch. 6).

According to Haidt, the clustering of moral rules around certain themes is the result of innate predispositions that facilitate the learning of certain norms. In

---

5 In some other works Haidt calls this domain "Loyalty/Betrayal Foundation" (Haidt, 2012).
psychology, it is universally accepted that some things are easier to learn than others. It is extremely difficult to mold a child’s mind when the effort is made in the opposite direction to what she likes naturally. It does not require much effort, for example, to make a child prefer candies to broccoli, the sympathy of other children rather than the approval of adults, or to retaliate aggression instead of loving her enemies (Haidt & Bjorklund, 2008, p. 201). The central idea of this kind of model is that humans have strong predispositions to develop certain reactions and preferences (Giroux, 2011, p. 292-3). What we consider morally permissible, forbidden or obligatory may be a result of these innate emotional propensities combined with our cultural experiences, without the need to apply a complex set of principles, such as those involved in language.

The hypothesis that human beings have a predisposition (preparedness) to moral learning is reinforced by experiments that show the early development of perceptions which are involved in moral judgments. Before walking, children are already able to recognize and value behaviors such as helping or harming others. Hamlin, Wynn & Bloom (2007) presented to children aged between 6 and 10 months some performances in which a puppet was trying to climb a slope. In some of the presentations, the puppet was helped by another puppet that pushed it up. In others, a third puppet appeared on the slope and struck the rising puppet, preventing it from reaching the top. After these presentations, both puppets were put in front of the children, who showed a strong preference for the one who tried to help. According to Haidt and Kesebir, experiments like this indicate the existence of a perception system capable of creating positive emotions towards helpers, and negative emotions against bullies (Haidt & Kesebir, 2010, p. 804). Emotions like these, in turn, could facilitate the adoption of rules designed to avoid actions that harm third parties.

In turn, this emphasis on the role of emotions brings us to the affective resonance model elaborated by Shaun Nichols. Inspired by Dan Sperber, he advocates an epidemiological model that emphasizes how emotional dispositions can restrict moral development possibilities. The hypothesis formulated by Nichols predicts that, all other circumstances kept unchanged, rules prohibiting actions that have a high probability of awakening negative emotions have a higher chance of being assimilated and transmitted than rules that are not connected to emotions (Nichols, 2008, p. 270). From Nichols’ perspective, our emotional dispositions influence our moral judgments,
but should not be confused with them. He tries to explain how emotions and norms interact. To this end, he resorts to the major role cultural evolution played in determining which norms have 'survived' throughout history. Those rules prohibiting actions that are likely to produce negative emotions have more cultural fitness. He acknowledges there are numerous factors that influence cultural evolution and affective resonance is just one of those: emotionally conspicuous cultural aspects tend to attract our attention and to be memorized, and are, therefore, more likely to endure (Nichols, 2008, p. 269-70).

According to Nichols’ hypothesis, rules designed to prevent the production of damage and injury (harm norms) have an advantage over other rules in the process of cultural evolution since ‘normal’ human beings have a strong aversive reaction to suffering. As well as other basic emotions such as sadness, anger, disgust and fear, emotional reactions to the suffering of others are believed to be universal and innately specified. As a result, in all cultures, actions that cause suffering to others tend to arouse aversion (Nichols, 2008, p. 271). Rules to prevent harm could have emerged for different reasons at different times, but the important point is that when this happens they find an important ally in emotions.

Thus Nichols attempts to explain the existence of certain moral norms as a result of innate biases. Yet, he stresses these biases are part of an affective innate system, which means they are neither information in a propositional form nor a set of innate moral principles (Nichols, 2005, p. 369). For Nichols, the moral domains originate from emotional systems, especially the affective system that responds to suffering in others. If this type of model is correct, emotions can play a role in shaping cognitive structures, reducing the appeal of theories that propose the existence of innate propositional knowledge (Nichols, 2005, p. 368).

---

6 For this reason, experiences with children who show aversion to harmful practices should be interpreted with caution. The fact that a baby expresses aversion to certain offenses, does not mean that she judges that this behavior is morally wrong.

7 Affective resonance hypothesis: “Norms that prohibit actions to which we are predisposed to be emotionally averse will enjoy enhanced cultural fitness over other norms.” (Nichols, 2008, p. 269-70)

8 Sripada offers a similar explanation. He argues that changes in moral norms would be better explained by what he calls thematic clustering. In virtually all human groups, morality deals with some common issues. Sripada says that the existence of this thematic clustering is the result of innate biases acting on the content of moral norms. The central idea of this innate biases model is that there are innate structures favoring the arousal of a feeling of aversion or sympathy towards certain behaviors and, consequently, favoring the emergence and maintenance of certain moral norms (Sripada, 2008, p. 330-337).
Although Nichols emphasizes Haidt's harm/care foundation, he acknowledges the existence of other candidates for the role of moral universals. He admits the importance of emotions associated with fairness and, to a lesser extent, disgust in shaping morality (Nichols, 2005, p 356; Nichols, 2008, p 266). The history of etiquette norms, for example, demonstrates how norms associated with the emotion of disgust have a greater chance of remaining unchanged (Nichols, 2002).

The central aspect of the models developed by Haidt and Nichols is the relevance attributed to innate dispositions, especially certain types of emotions, in shaping the morality of a given human group. According to them, there is not an innate moral knowledge, as stated in UMG, but a tendency to moralize certain behaviors based on the emotions they arouse.

3 Shortcomings of the Moral Foundations approach

Jesse Prinz is one of the main critics of moral nativism. In general, he attempts to demonstrate that the capacities involved in moral judgments are not adaptations specific to morality. However, this kind of argument is relevant to the refutation of moral nativism only when it presupposes an evolutionary concept of innate character.

We can group, following the classification proposed by Joyce, the definitions of innateness into two main groups: the evolutionary conception and the developmental conception. According to the first definition, moral nativism is equivalent to the claim that morality is an adaptation in the Darwinian sense, that is, was selected for by means of natural selection (Joyce, 2013, p. 532). On the other hand, those who adopt a developmental conception consider that a trait is innate when its appearance is protected from variations in the environment where the development of the individual occurs (Joyce, 2013, p. 533).

Haidt seeks to explain from an evolutionary perspective the emergence of emotions related to morality and he certainly believes that these emotions were selected for (they would be, in this case, adaptations). Perhaps he believes that the ability to make moral judgments is also an adaptation, but this is not the most important aspect of his kind of moral nativism. The models of Haidt and Nichols hold that innate characteristics, in a developmental sense (either adaptations or side effects of adaptations), ensure that morality develops and retains certain characteristics even in different environments. Therefore, the criticism of Prinz does not reach the most relevant aspects of the models developed by Haidt and Nichols. Nevertheless, there are...
two criticisms that seem more relevant in relation to the MFT: the fact that it is committed to a modular vision of the human mind and its incompleteness, since it does not provide an explanation for the mental steps previous to the triggering of intuitions.

Haidt adopted in his MFT the idea that the human mind is modular, but rejected some of the features that Fodor (1983) used to characterize modularity. He adopted a model inspired in Sperber's massive modularity hypothesis (2005): there is a set of innate learning modules capable of producing more specific modules during the development of the individual (Haidt & Joseph 2007, p. 379-80). In other words, there is an innate learning module – first-order module - for each moral domain and each of these modules produces, from the experiences of the individual, several other working modules responsible for generating the moral intuitions in specific situations - second-order modules⁹.

The moral dumbfounding phenomenon is the main reason why Haidt believes the human mind is modular. In many cases, people keep their moral judgments unchanged despite being confronted with new information that invalidates the justifications used to support the original judgments. Haidt thinks the explanation for this is that the systems that produce these judgments are encapsulated to some extent. Unlike Fodor, Haidt admits that the modules have access to information stored elsewhere in the mind, but in a limited way. Haidt also holds that the modules dedicated to morality belong to specific domains.

Nonetheless, it is not clear how Haidt came to the conclusion that each of the five moral foundations corresponds to a different mental module (Giroux, 2011, p. 294-5). The simple fact that we are able to classify intuitions about morality in five categories (foundations) does not justify the belief that each of them is implemented independently or by a discrete computational mechanism (Mallon, 2008, p. 151). The foundations proposed by Haidt could be universal and have an innate basis, but that does not mean there are five specialized modules (Giroux, 2011, p. 294-5). While it is right that some things are easier to learn than others, this fact does not justify the inference that the ease of learning results from the existence of learning modules dedicated to specific domains.

The idea that the human mind is composed of several specialized mechanisms dedicated to solving particular types of problems faces strong opposition. Even Sperber

⁹ We took this expression – “second-order modules” – from Suhler & Churchland (2011, p. 2104).
recognizes that only a small number of cognitive scientists believe that the mind is massively modular. For most of them, the mind is largely non-modular. Although many admit the existence of modules related to perception, just a few argue that the central systems that process these inputs are modular (Sperber, 2005, p. 53). In addition, the model of modularity proposed by Sperber and adopted by Haidt is not consilient\(^\text{10}\) with the empirical results of neuroscientific research. If the human mind were made up of modules, we should expect the organization of the brain to be at least compatible with the existence of these modules (Suhler & Churchland, 2011, p. 2109). However, the anatomy of our central nervous system makes the ideas of informational encapsulation and domain specificity implausible.

Local neural connections in the cortex are dense, while connections that span greater distances are more sparse, a fact known as small world architecture. Despite this, just a few synapses separate a particular neuron from any other present in our brain. As stated by Suhler and Churchland, everything is easily accessible to everything else in a few synaptic steps (2011, p. 2109). This pattern prevails even in the primary visual cortex (V1), the area responsible for receiving inputs from the retina through the Lateral Geniculate Nucleus (LGN). Over 80% of synaptic contacts present in V1 do not come from the LGN, but from other brain regions (Suhler & Churchland, 2011, p. 2109). As a result, the V1 operation depends on several other factors.

The assumption that modules are responsible for specific domains also confronts similar challenges. Prinz demonstrates that the usually mentioned instances of modules do not correspond to specific domains. He points out, e.g., how mind-reading relies on working memory (2006, p. 28). Prinz cites research conducted by Mckinnon and Moscovitch which showed that the performance of individuals whose working memory was kept busy was impaired in tasks related to the attribution of beliefs (McKinnon & Moscovitch, 2007). Prinz further highlights how neuroimaging studies show that mind-reading involves the use of several brain regions, and that each of these regions also contributes to many other capacities, i.e., they are not specific to any domain. Likewise, moral judgments recruit various areas of the brain responsible for different capacities, including those normally associated with emotional centers (2006, p. 29-30).

\(^{10}\) We use the word ‘consilient’ in the sense proposed by Whewell: “The Consilience of Inductions takes place when an Induction, obtained from one class of facts, coincides with an Induction, obtained from another different class. This Consilience is a test of the truth of the Theory in which it occurs” (2012, pos. 7172).
Haidt is aware of the objections directed towards the modularity of mind. As a result, he argues that the existence of modules is not a central aspect of his MFT. The foundations he proposes could be explained more generally in terms of preparedness. A milder version of MFT could be described as follows: the human mind has been shaped by evolutionary processes in such a way that children learn easily to worry about the avoidance of harm, the correction of injustices (fairness), the preservation of the members of their own groups (in-group), the respect for authority and about purity issues. Notwithstanding, this does not mean they have any innate moral knowledge; it only means they are prepared to acquire certain moral beliefs and to resist others (Haidt & Bjorklund, 2008, p. 204).

When reformulated this way, Haidt's MFT becomes akin to the epidemiological model proposed by Nichols. According to Nichols, it is plausible that the emotional systems dedicated to reacting to the suffering of others have evolved as a way to overcome environmental challenges existent during the Pleistocene. Furthermore, their influence on cognitive structures is not restricted to specific areas. The emotions that affect moral judgments are not specific to this kind of appraisal, they can affect the acquisition of knowledge in other areas. Our responses to suffering could, e.g., affect the way we think about natural disasters that cause human misery (Nichols, 2005, p. 368). This pervasive influence of emotions on behavior, inasmuch as it indicates the absence of encapsulation, challenges the idea the mind has a module or group of modules specific to morality.

Nichols’ affective resonance model and what is left of MFT essentially claim that (1) we learn some things easier than others and (2) our emotions play an important role in determining what will be easier to learn. These statements are not very controversial because it is hard to imagine how dispositions such as our basic emotions - sadness, anger, disgust, fear etc. - could not interfere with learning and the development of social norms (Giroux, 2011, p 289.). However, this kind of claim is not enough to clarify many of our questions about moral nativism: Haidt says the human mind is prepared to learn certain things, but does not provide details about what exactly this innate organization that favors learning comes to. He does not explain (at the level of cognitive psychology, developmental psychology, neuroscience, etc.) how humans are prepared to acquire moral norms (Suhler & Churchland, 2011, p. 2105). Similarly, Nichols's claim that our emotions favor the adoption of certain norms needs a more
thorough explanation. Haidt and Nichols did not unveil the details of how cognitive processes result in the production of moral judgments.

Another major criticism of the MFT relates to the fact it does not give due attention to cognitive elements prior to the outbreak of emotions and intuitions. A creature can have an emotion only after its mind identifies that the situation is worthy of that emotion. There is a previous unconscious analysis responsible for identifying the causes and consequences of an action - who did what? why? by which means? to achieve which goals? - and triggering an emotional reaction (Hauser, 2007, p. 8). Hauser (2007) and Mikhail (2011) point out that very similar actions cause completely different emotional reactions depending on how they are perceived or mentally represented by those who watch them. Depending, for example, on the intentions we attribute to the author (mindreading), we will have different reactions. Emotions alone could not explain why an action is considered wrong in one context, but correct in another. Therefore, we have to understand how mental representations, emotions and other slower cognitive processes interact. A full explanation of how moral judgments occur should clarify the computational structures responsible for analyzing the scenario where the action to be judged took place (Dwyer, Huebner & Hauser, 2010, p 494).

Moral judgments rely on an interpretation of the action made by the person who judges. When we make these judgments we use a number of inferences about the intentions and mental states of the author of the action. The abilities required to make these inferences are not restricted to the realm of morality. We employ them when we interpret any action, no matter how irrelevant it might be to morality. The distinction between intentional and accidental actions, e.g., usually is relevant to moral judgments, as many actions are only considered reprehensible when performed intentionally. Notwithstanding, when we observe a child playing with a ball, we are able to distinguish if she let the ball slip from her hands or if she bounced the ball on purpose, although it does not have any moral relevance. Similarly, when we see a child cutting a lemon, we assign a goal to that act, e.g., making lemonade. We constantly infer invisible properties of the mind through indirect verification, e.g., what someone else is looking at, which objects this person is trying to reach or where she came from. A theory of mind is needed to understand the intent and the action of the aforementioned child, but it does not imply a moral judgment. This capacity is like the capacities for visual perception or memory: they can be used in moral judgments, but are not specific
to that sphere, i.e., they can also be used in a range of other tasks (Sripada, 2008b, p 362; Hauser, 2007, p. 50-1).

The way we interpret an action can provoke a feeling of aversion or attraction. This feeling motivates the adoption of a certain reaction and explains in large part how we arrive at a moral judgment. However, the simple reaction (aversion or attraction) to a particular circumstance does not depend on the existence of a moral judgment. Even an amoeba is able to identify certain features of its environment – concentration of ions, for instance - and react, moving closer or moving away. Obviously the amoeba’s reaction does not call for the application of a set of principles and parameters with a complexity near to that of a grammar. When we face a particular environment or action we also react according to the characteristics we identify. Nonetheless, the simple emotional reaction (aversion or attraction) to a particular conduct does not fully explain our moral judgment, since this involves a diverse cognitive element, a belief about the value of the performed action.

The distinction between a simple emotional reaction and a belief is useful to demonstrate that morality is not a monolithic entity. That is, it involves a number of elements. As we shall see, realizing this is crucial to answering the fundamental question of the debate on moral nativism.

4 So, is morality innate?

If we adopt a developmental concept of innateness, possibly we will conclude that morality is innate. It is present in every community we know of and, in general, the existing moral norms in these different communities cluster around the same themes. These facts suggest that there is a biological basis responsible for producing certain biases in moral learning. Many opponents of moral nativism agree that the acquirement of norms is biologically prepared. However, they believe that the crucial adaptations are perceptual and motivational and that they are not specific to morality. In other words, they adopt an evolutionary concept of innateness. If they adopted a developmental concept they would conclude that morality, or at least most of its elements, is innate. Prinz, for example, considers morality a side effect of psychological traits that have

---

11 Nichols recognizes that morality is not a monolithic entity. He believes that it is a side effect of at least two other features: “both of the mechanisms that I’ve suggested contribute to moral judgment might well be adaptations. However, it is distinctly less plausible that the capacity for core moral judgment itself is an adaptation. It’s more likely that core moral judgment emerges as a kind of byproduct of (interalia) the innate affective and innate rule comprehension mechanisms” (2005, p. 369).
evolved for other purposes. He recognizes that morality is constrained by our biological characteristics and that we are not born as a tabula rasa. According to Prinz, our emotions, our ability to attribute mental states and the care we have for our relatives function as building blocks of morality, but should not be confused with it (2013).

On the other hand, if we adopt an evolutionary concept of innateness, the issue becomes more complicated. It is likely that many of the aspects related to what Darwin called *social instincts* are adaptations. That is, all those emotions that lead us to cooperate within our societies and that end up producing what Haidt calls *moral foundations* were probably selected for because they favored certain behaviors. However, morality cannot be reduced to *social instincts*, as it relies also on what Darwin called *intellectual powers*12.

The distinction made by Darwin between *social instincts* and *intellectual powers* is still relevant, since it makes clear that the existence of groups of altruists who do not have a moral faculty is at least conceivable (Joyce, 2006). For the animals to behave in one way or another they do not have to be able to judge any behavior as good or bad: the selection of a behavior does not require that it had been consciously adopted. That is, there could be some kind of being who behaved in a way we consider morally praiseworthy, but was incapable of making any moral judgment. We could act altruistically by some inclination, without having the belief that we ought to act in this way (Prinz, 2013, p. 107). Doing something because we like it is different from doing something because we believe we have a duty (Joyce, 2006, p. 50).

Self-consciousness is one of those intellectual capacities indispensable for morality. In order that we can judge normatively an action that we ourselves practiced, we must be able to identify the reasons that led us to practice it and to reflect on these reasons. That is, we must be able to compare them to other reasons that could have motivated us to act differently. Thinking in a normative way requires being aware of the grounds of our beliefs and actions, and this implies self-consciousness because it involves the capacity to identify ourselves as the subjects of our mental representations. A being without self-consciousness can be aware of the existence of an object that she wants and act upon this information. On the other hand, a being with self-consciousness is also aware of the fact that *she* wants the object. She does not think just about the object she wants, but also about her own desires that make her willing to act in certain

---

12 Darwin believed that the emergence of morality was inevitable whenever an animal had strong social instincts combined with an intellectual capacity as developed as that of humans.
way. This self-consciousness about the motives ensures a reflective distance that allows the subject to question her own motives (Korsgaard, 2006, p. 112-116). Through self-consciousness the individual is able to assume the position of a spectator of her own desires. From that position, she can compare her past and future actions and approve or disapprove of them. She may think it would have been better if she had acted according to another desire (Darwin, 1871, p. 73-74).

This kind of thinking requires other skills such as memory and language. The capacity to reflect on the motives that led to an action depends on the capacity to remember the action, its motives and its results. The memory that an action provoked a feeling of dissatisfaction can, e.g., provide subsidies for the judgment that we ought to have acted otherwise (Darwin, 1871, p. 70-72). Language, on the other hand, besides allowing conscious reflection on the grounds of our actions and the formulation of the belief that we ought to have done otherwise, also allows us to share these reflections and beliefs with other members of our group. As shown in the social intuitionist model developed by Haidt (2012), this kind of social interaction is another important element in the formation of morality.

Throughout this work, various capacities involved in the moral judgments have been discussed: pattern recognition, mind-reading, self-consciousness, language, memory, emotions, etc. Each of these capacities is used in the formulation of moral judgments, but is not dedicated solely to this task. Thus, it is quite plausible that the ability to make moral judgments has arisen as a secondary effect of other mental capacities. Nevertheless, this does not mean that the capacities involved in moral judgments have not been modified as a result of their contribution to these judgments. At some moment, morality may have worked as an exaptation, *i.e.*, as a feature that enhances fitness, but that has not evolved as a result of selective pressures related to its current role. The capacity to make moral judgments exists in different degrees depending on the possible combinations of the elements that constitute this capacity. Certain compositions of these elements are more adaptive than others and can be selected. Thus, even if morality has emerged as a side effect, it might have undergone subsequent adaptive structural changes, possibly because of interactions with cultural changes. Morality can be a *secondary adaptation*, as long as its elements have been modified by natural selection in consequence of the fact that the ways they interact have an impact on the fitness of individuals or groups (Joyce, 2014, p. 127-128).
Realizing that morality involves all elements previously mentioned makes clear the difficulty of discerning if it is an adaptation or just a *spandrel*, and also illustrates the shortcomings of the main approaches adopted so far.

On the one hand we have the theory of universal moral grammar, which aims to explain some details of moral judgments, but fails to provide a plausible explanation because of the major differences between the processes involved in morality and in the language faculty. On the other hand, we have the moral foundation theory, which, in its mild version, highlights the influence of emotions on our moral judgments, but fails to provide a detailed account of many elements required for the development of morality.

In this sense, neither the evolutionary, nor the developmental theory can offer a complete explanation of morality. From the developmental perspective, in order to explain that which is innate in morality, we would have to detail the capacities involved, how they interact and how they distinguish themselves from proximate cultural causes. On the other hand, responding as if morality is an adaptation also involves the investigation of how and when these capabilities have emerged and which selective pressures have acted in this process. In this case, we should still investigate if these pressures were sufficient to modify, by natural selection, the capacity to make moral judgments. Up to now, the authors who have studied the evolution of morality have not done much more than an investigation about the evolution of cooperation. However, a full account of the evolution of morality should also involve a story about the evolution of consciousness (or self-consciousness) as a condition for self-control. By means of self-consciousness, the agent becomes able to decide whether or not to adopt a purpose, judging it as good or bad, subject to approval or not.

**References**


SPERBER, Dan. Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence, & S. Stich (Eds.),