Universal moral grammar: a critical appraisal
Pierre Jacob, Emmanuel Dupoux

To cite this version:
Universal moral grammar: a critical appraisal

Emmanuel Dupoux\textsuperscript{1,2} and Pierre Jacob\textsuperscript{3}

A new framework for the study of the human moral faculty is currently receiving much attention: the so-called ‘universal moral grammar’ framework. It is based on an intriguing analogy, first pointed out by Rawls, between the study of the human moral sense and Chomsky’s research program into the human language faculty. To assess UMG, we ask: is moral competence modular? Does it have an underlying hierarchical grammatical structure? Does moral diversity rest on culture-dependent parameters? We review the evidence and argue that formal grammatical concepts are of limited value for the study of moral judgments, moral development and moral diversity.

Introduction
The task of moral psychology is to provide an account of the mechanisms underlying human moral judgments. A new framework for moral psychology is currently receiving much attention: the so-called ‘universal moral grammar’ (UMG) [1–3]. First considered by Rawls [4], it relies on an intriguing analogy between the study of moral cognition and Chomsky’s linguistics research program called generative linguistics [5,6]. Given the impressive achievements of generative linguistics in the investigation of the human language faculty (HLF), UMG deserves serious consideration. Here we assess UMG and argue that the formal grammatical concepts proposed by Chomsky are unlikely to throw much light onto the psychological mechanisms that underlie moral judgments, moral development and moral diversity.

Chomsky’s revolution in linguistics
The starting point of generative linguistics is the fact that speakers of any human language are able to understand and produce novel sentences that they have never perceived before. Furthermore, it rests on three major premises.

First, Chomsky [5,6] argued that a speaker’s ability to understand and produce novel sentences depends on her tacit knowledge of a grammatical system, of which the speaker is totally unaware. Generative linguistics investigates the computational properties of this deeply unconscious grammatical system. Its task, therefore, is sharply demarcated from what Chomsky [7] calls the ‘ethno-scientific’ study of a speaker’s explicit (folk or naïve) metalinguistic beliefs about his or her language.

Second, the grammar of a language is a finite system of recursive rules whereby an unbounded set of sentences can be generated from a finite lexicon (Figure 1a). As Chomsky pointed out, the evidence shows that this grammatical system is dedicated to language processing and encapsulated from explicit beliefs: a speaker’s grammatical competence is a modular cognitive system [6,7].

Third, all human children manage to acquire knowledge of their language from a finite and fragmentary sample. From this phenomenon, called the ‘poverty of the stimulus’, Chomsky argued that the initial state of the HLF must include innate knowledge of a so-called ‘universal grammar’ (UG), that is, a set of universal computational principles that guide a child towards the appropriate grammatical generalizations [7]. Thus, Chomsky’s nativism grounds the study of the HLF into human biology.

This research program has given rise to the so-called ‘principles-and-parameters’ approach to linguistic diversity, according to which UG (the initial state of the HLF) is conceived as a fixed, species-specific and task-specific (i.e. modular) network of computational principles connected to a finite set of binary parameters, each of which can occupy one of two positions (e.g. either ‘head first’ or ‘head last’). Each human language turns out to instantiate a particular setting of the parameters [7,8].

Universal moral grammar: a fascinating prospect
Given the achievements of generative linguistics in addressing language acquisition and linguistic diversity, addressing moral development and the diversity of moral belief systems from the standpoint of UMG, that is, a moral psychological framework based on the analogy between the human moral faculty and the HLF, seems to be a promising strategy. As advocates of UMG point out [1–3], humans are able to assess the moral properties of an unbounded set of social interactions, and UMG can be usefully organized around five main questions, each of which has a counterpart in generative linguistics: (i) What is the structure of human moral competence? (ii) How is it acquired by a human child? (iii) How is it put to use in human action?
(iv) How is it realized in the human brain? (v) How did it evolve in the human species?

Three results from the empirical study of moral judgment seem to support the UMG program for research into the moral faculty. First, as exemplified by what Haidt [9,10] has called ‘moral dumbfounding’, people often exhibit strong moral convictions for which they cannot offer explicit justifications (Box 1). Similarly, people are unable to justify their grammatical judgments. Secondly, as Turiel [11] has argued, unlike social conventional rules, which depend on the acceptance of some social authority and are revisable at the authority’s request, moral norms are not revisable at some authority’s request (Box 1). In this respect, moral judgments are like grammatical judgments.

Finally, examination of so-called ‘trolley dilemmas’ suggests that moral judgments might, like grammatical judgments, depend on complex and unconscious computations involving abstract hierarchical structures. For example, examination of the minimal contrast between the trolley dilemmas, respectively called Loop Track and Man-in-Front (Box 1), shows that moral judgments depend on whether some consequence of an action was intended or merely foreseen by the agent [2,3,12–14].

None of this evidence, however, shows that the mechanisms underlying moral judgments make use of moral information encoded in a dedicated moral grammar. To probe the UMG research framework, we first address the issue of whether moral competence is a modular system separable from an agent’s explicit moral beliefs. Second, we examine whether moral competence is best conceived as a recursive grammatical system of computations. Third, we examine the idea that moral diversity can be derived from a system of universal principles and binary parameters.

How modular is moral competence?

The top-down modulation of moral judgments

One piece of evidence for the modularity of grammatical competence is that speakers’ grammatical judgments are independent of their explicit metalinguistic beliefs about their language. The discovery of moral dumbfounding has contributed to the emergence of so-called ‘dual process’ models of moral judgment [10,15–21]. On these models, social intuitions, which subsequently serve as inputs to moral judgments, are automatically generated in response to the perception of human interactions (Figure 1b). For example, Haidt’s [10,18,19] ‘social intuitionist’ model is based on the separation between the processes underlying respectively moral judgments and explicit justifications. Arguably, moral justifications are meta-cognitive processes operating on explicit moral beliefs.

However, processes of moral justification have been scrutinized for centuries by moral philosophers from Plato and Aristotle to anthropologists and were later investigated by Piaget [22] and Kohlberg [23,24], and it would be hard to claim that these processes do not belong to the human moral faculty. Furthermore, moral dumbfounding...
**Box 1. The seven pieces of the moral puzzle**

**Moral dumbfounding**
A brother and sister intend to make love with one another once and only once in their life. They decide to keep it a secret between themselves, and to take all protective measures against pregnancy, and they both have a wonderful sensuous experience. Are the siblings’ actions OK?

Most, if not all, of the subjects who judge the action morally objectionable feel dumbfounded and find themselves utterly unable to explain why [10].

**Non-revisability at the request of a social authority**
(i) Tom wears pajamas to go to school. Is it OK?
(ii) Tom hits the little girl next to him for no reason. Is it OK?

Suppose that the teacher said it was OK, would it now be OK to do (i) or (ii)?

Adults and young children judge that, unlike violations of conventions (i), violations of moral norms (ii) are wrong, even if an authority decides it is OK [11,41,42].

**The intentional complexity of trolley dilemmas**

**Bystander**
A train is about to run over and kill five people walking on the tracks. A bystander can flip a switch that will divert the train onto a sidetrack, and save five but kill one person who is standing on the sidetrack. Is it permissible to flip the switch? [14]

**Footbridge**
A train is about to run over and kill five people walking on the tracks. A bystander standing on a footbridge can shove a large man onto the tracks and thereby save five but kill the large man. Is it permissible to shove the man? [14]

**Loop Track**
A train will run over five people. By flipping a switch, a bystander can temporarily divert the train onto a sidetrack loop which goes back to the main track, where the five people are standing. On the sidetrack there is one person who will be killed by the train, thereby stopping the train and saving five people. Is it permissible to flip the switch? [14]

**Man-in-Front**
Same as Loop Track, except that on the sidetrack there is a heavy stone that will stop the train, thereby saving five people but killing one person standing in front of it. Is it permissible to flip the switch? (Mikhail, 2000, Rawls’ linguistic analogy, PhD dissertation, Cornell University)

Even though the casualties are the same, people judge that it is more permissible to sacrifice one person to save five in Bystander than in Footbridge and in Man-in-Front than in Loop track. In each pair, the relevant contrast is between causing one person’s death as an intended consequence or merely as a foreseen consequence of the agent’s act [2,3,12–14].

**The dissonance between intuitions and explicit moral beliefs**

- We are positively biased toward members of our in-group (relatives, kin, social or ethnic group) and negatively biased toward members of an out-group [27]. In pre-industrial societies, hostile behavior toward members of out-groups is often praised and morally sanctioned. A less explicit version of such biases is demonstrated in cultures where impartiality is highly praised [26]. Race-related biases can be revealed by implicit measures that fail to correlate with explicit racist or anti-racist beliefs [26].

- Survivors of catastrophes can experience a strong sense of guilt, even if, according to their reflective moral standpoint, they believe that they share no responsibility for the victims’ fate – this is the “survivor guilt” syndrome, part of the Post Traumatic Stress Disorder syndrome [43].

**The cognitive cost of moral dilemmas**

- A loved one is suffering from a disease for which there is no known cure and which causes her insufferable pain. Should you pay an end to her suffering by abbreviating her life?

- Alternatively, suppose that there is a cure for the disease, but the medicine is so expensive that you cannot afford it. Should you steal the medicine? [23,24]

Such dilemmas arise from the experience of conflicting intuitions (e.g. relieving pain versus avoiding stealing). Contrary to perceptual or linguistic ambiguities, the solution of such conflicts is not automatic but requires effortful strategic thinking.

**Empathetic modulation**

- Empathy with the victim
A person has just been unfair to you. You know that he receives an electric shock when this happens. Your empathy for him and your empathetic brain response correlate negatively with the degree to which you judge his pain to be well deserved [44].

- Empathy with the agent
In some cultures, when crimes of passion are committed under the influence of emotions such as jealousy, they are considered as more permissible and are less harshly punished than cold-blooded crimes. In other cultures, so-called “honor killings” are planned, premeditated and ritualized accomplished. In cultures of both types, members of the local community might empathize more with the agent’s emotions than with the victim’s, thereby modulating the perceived gravity of the crime (pp.142–155) [2].

**Lack of empathy in psychopaths**
Adults and children with psychopathic tendencies are impaired in the recognition of sad and/or fearful emotions and show no behavioral empathy and no empathetic brain responses to others’ pain or distress. They also fail to distinguish moral from conventional norms [30,32,45].

**Moral diversity**

- In different cultures, by and large, humans seem to exhibit the same automatic, encapsulated, emotional responses to social interactions [46]. Cultural anthropologists propose that moral systems are based on a small set of universal moral values (cf. Shweder’s threefold distinction between the ethics of autonomy, community and divinity [34,35]).

- Across different cultures, systems of justifications for moral decisions appeal to different supernatural intentional agents, such as gods, spirits or ancestors, posited by distinct religious belief systems. According to Boyer [49] (pp.120) and Atran [47] (pp.110–112), imagining empathetic support from a supernatural intentional agent may facilitate the adjudication of moral dilemmas and the justification of hard moral decisions [48].

- Shweder et al. [36] studied preferences for sleeping arrangements among Indian and North-American middle-class families with five children of both sexes and two parents. They found that, given constraints on the number of rooms, the choices of Indian families revealed the following preference ordering among moral values: (i) incest avoidance, (ii) protection of the vulnerable, (iii) female chastity anxiety. By contrast, the choices of middle-class North American families revealed the following preference ordering: (i) incest avoidance, (ii) sacred couple (parents’ privacy), (iii) children’s autonomy.

If an agent’s moral judgments depend on her explicit moral beliefs and preferences, then there can be no sharp demarcation between the scientific investigation of the human moral faculty and the ‘ethnoscientific’ study of people’s naive explicit moral beliefs. If not, then one major step toward the assumption that moral judgments derive from tacit knowledge of a moral grammar is missing.
The cognitive cost of moral dilemmas
A speaker’s intuitions about the grammatical properties of a sentence exhaust, and are constitutive of, that person’s grammaticality judgments (Figure 1a). They are generated quickly, automatically and effortlessly. However, the existence of moral dilemmas shows the possibility of a gap between one’s intuitions about the moral properties of an action and moral judgment. Moral dilemmas arise from a conflict between two or more deeply felt obligations pulling in opposite directions (Box 1). As emphasized by Greene [15–17,20,21] and other advocates of dual-process models, intuitions are based on distinct, fast, automatic, emotional mechanisms that react specifically to certain social situations. If and when conflicts arise among competing intuitions, the task of moral judgment is to adjudicate them (Figure 1b). Because it is open to top-down modulation, the process of moral adjudication can be time consuming (Box 2).

Moral dilemmas have no counterpart in the HLF. When a sentence of a natural language is syntactically or semantically ambiguous, it has two or more underlying grammatical structures. Because the ambiguity is automatically resolved by pragmatic mechanisms operating on contextual cues, ordinary speakers are rarely aware of it and it is unlikely to disrupt the continuous flow of processing [28]. If a speaker does become aware of it, then contrary to the case of moral dilemmas, the person is unlikely to feel torn by the experience of conflicting intuitions. In sum, moral competence does not seem to be modular in the relevant sense.

Is there a moral grammar?
We now argue that moral competence exhibits a feature that grammars of natural languages lack and that it fails to meet the requirements for having a grammatical structure.

Emotions and moral competence
The idea that a speaker’s linguistic competence has a grammatical structure would be hard to reconcile with the claim that emotions play a causal role in the generation of grammatical intuitions. This is why Hauser [2,29] argues that emotions belong to moral performance, not competence. In his view, moral judgments cause emotions, but not vice versa. Although the issue is not settled yet, there is evidence that emotional responses contribute to moral competence: not only do violations of moral norms automatically elicit emotional responses, but as Blair [30–32] has argued, the lack of empathy with the distress cues of victims interferes with normal moral development in children with early psychopathic tendencies (Box 1 and Box 2).

Four conditions of adequacy for moral grammar
The basic task of a generative grammar is to characterize the recursive mapping between pairs of phonological and semantic representations for an infinite number of such pairs from a given language. The mapping has at least four features. First, the phonological and semantic properties of a sentence are a function of the phonological and semantic properties of their constituents. The grammatical (or syntactic) mapping must therefore reflect the compositionality of the phonological and semantic representations. Second, because language is used as a communication system, the mapping preserves information and is reversible: phonological representations can be mapped onto semantic representations (comprehension) and semantic representations can be mapped onto phonological representations (production). Third, the information used in the psychological processes implementing this mapping is encapsulated in Fodor’s [33] sense, in that it depends on linguistic information alone, not on general world knowledge. Fourth, the grammar is domain-specific in that the linguistic rules are unique to the HLF.

What could the computational task of MG be? Suppose that MG maps the structural description of an agent’s act onto its moral valence [3]. If so, then the question arises: could the mapping satisfy the compositionality, reversibility, informational encapsulation and domain-specificity requirements?

A complex action can be conceived as a function of its constituents. But for two related reasons, the valence assigned to an agent’s act cannot be a function of the valences associated with its constituents. First, the valence of an act lacks discrete constituents or a hierarchical structure. Second, the component parts of a morally valued complex act (e.g. opening a bottle, pouring the contents in a tea cup, serving tea to a guest), might in themselves carry no negative valence at all. Only if one knew that the bottle contained poison would one assign a negative valence to the act.

The mapping is not reversible; it makes sense to map the structural description of an agent’s act onto a valence. But it would make little (if any) sense to generate a structural description of an act in response to the representation of the valence of that act. Morality is an evaluative system, not a generative one.

The psychological processes implementing this hypothetical mapping are unlikely to satisfy a strict informational encapsulation requirement; any background information about an action (e.g. knowledge of what a poison is, the past actions of the agent and the victim, etc.) might affect moral evaluation (see also Box 1).

Could MG be domain specific? We discuss two possible arguments. First, it might seem as if Mikhail’s [3] computational analysis of the Bystander example shows that the structural description of an agent’s act can only be mapped onto the valence of the agent by a domain-specific MG, but it does not. What Mikhail calls the
‘moral structure’ of the action is purported to be mechanically derivable by so-called ‘moral transformations’ from the underlying ‘causal structure’ of the action. Each of these ‘causal structures’ turns out to be the semantic structure of an English sentence describing the action (step [d], Figure 2 of Mikhail [3]). The alleged moral transformations turn out to be rewritten rules that insert a specific moral predicate into these semantic structures (step [e]). Far from showing that our intuitions about the moral properties of actions are generated by a dedicated MG, the analysis shows that the grammars of natural languages have the resources to encode information about some of an action’s properties that are relevant to moral judgment.

Second, it might seem as if the minimal pair of trolley dilemmas, called Loop Track and the Man-in-Front, respectively (Box 1), reveals the domain specificity of the dilemmas, called Loop Track and the Man-in-Front, judgment. Some of an action’s properties that are relevant to moral languages have the resources to encode information about the ‘moral structure’ of the action is purported to be mechanically derivable by so-called ‘moral transformations’ from the underlying ‘causal structure’ of the action. Each of these ‘causal structures’ turns out to be the semantic structure of an English sentence describing the action (step [d], Figure 2 of Mikhail [3]). The alleged moral transformations turn out to be rewritten rules that insert a specific moral predicate into these semantic structures (step [e]). Far from showing that our intuitions about the moral properties of actions are generated by a dedicated MG, the analysis shows that the grammars of natural languages have the resources to encode information about some of an action’s properties that are relevant to moral judgment.

Second, it might seem as if the minimal pair of trolley dilemmas, called Loop Track and the Man-in-Front, respectively (Box 1), reveals the domain specificity of the dilemmas, called Loop Track and the Man-in-Front, judgment. Some of an action’s properties that are relevant to moral languages have the resources to encode information about the ‘moral structure’ of the action is purported to be mechanically derivable by so-called ‘moral transformations’ from the underlying ‘causal structure’ of the action. Each of these ‘causal structures’ turns out to be the semantic structure of an English sentence describing the action (step [d], Figure 2 of Mikhail [3]). The alleged moral transformations turn out to be rewritten rules that insert a specific moral predicate into these semantic structures (step [e]). Far from showing that our intuitions about the moral properties of actions are generated by a dedicated MG, the analysis shows that the grammars of natural languages have the resources to encode information about some of an action’s properties that are relevant to moral judgment.

Second, it might seem as if the minimal pair of trolley dilemmas, called Loop Track and the Man-in-Front, respectively (Box 1), reveals the domain specificity of the dilemmas, called Loop Track and the Man-in-Front, judgment. Some of an action’s properties that are relevant to moral languages have the resources to encode information about the ‘moral structure’ of the action is purported to be mechanically derivable by so-called ‘moral transformations’ from the underlying ‘causal structure’ of the action. Each of these ‘causal structures’ turns out to be the semantic structure of an English sentence describing the action (step [d], Figure 2 of Mikhail [3]). The alleged moral transformations turn out to be rewritten rules that insert a specific moral predicate into these semantic structures (step [e]). Far from showing that our intuitions about the moral properties of actions are generated by a dedicated MG, the analysis shows that the grammars of natural languages have the resources to encode information about some of an action’s properties that are relevant to moral judgment.

Second, it might seem as if the minimal pair of trolley dilemmas, called Loop Track and the Man-in-Front, respectively (Box 1), reveals the domain specificity of the dilemmas, called Loop Track and the Man-in-Front, judgment. Some of an action’s properties that are relevant to moral languages have the resources to encode information about the ‘moral structure’ of the action is purported to be mechanically derivable by so-called ‘moral transformations’ from the underlying ‘causal structure’ of the action. Each of these ‘causal structures’ turns out to be the semantic structure of an English sentence describing the action (step [d], Figure 2 of Mikhail [3]). The alleged moral transformations turn out to be rewritten rules that insert a specific moral predicate into these semantic structures (step [e]). Far from showing that our intuitions about the moral properties of actions are generated by a dedicated MG, the analysis shows that the grammars of natural languages have the resources to encode information about some of an action’s properties that are relevant to moral judgment.
nativist basis for moral competence. As we argued, moral judgments emerge from automatic emotional responses and complex computations tracking the intentional structure of social interactions. Rorty’s contention would only be corroborated if there were evidence that human emotions and human mind reading are not grounded in human biology. We are aware of no such evidence.

Acknowledgements
We thank Nicolas Baumard, Ned Block, Susan Carey, Hugo Mercier, Paul Smolensky, Dan Sperber, Dominique Sportiche, Steven Stich and Iris Trinkler for very useful discussion and comments.

References
4 Rorty, R. (2006) The influence of criticism on the contemporary German philosophical scene. In Acknowledgements, we thank Nicolas Baumard, Ned Block, Susan Carey, Hugo Mercier, Paul Smolensky, Dan Sperber, Dominique Sportiche, Steven Stich and Iris Trinkler for very useful discussion and comments.

References

Acknowledgements
We thank Nicolas Baumard, Ned Block, Susan Carey, Hugo Mercier, Paul Smolensky, Dan Sperber, Dominique Sportiche, Steven Stich and Iris Trinkler for very useful discussion and comments.

References

Acknowledgements
We thank Nicolas Baumard, Ned Block, Susan Carey, Hugo Mercier, Paul Smolensky, Dan Sperber, Dominique Sportiche, Steven Stich and Iris Trinkler for very useful discussion and comments.

References

Acknowledgements
We thank Nicolas Baumard, Ned Block, Susan Carey, Hugo Mercier, Paul Smolensky, Dan Sperber, Dominique Sportiche, Steven Stich and Iris Trinkler for very useful discussion and comments.

References

Acknowledgements
We thank Nicolas Baumard, Ned Block, Susan Carey, Hugo Mercier, Paul Smolensky, Dan Sperber, Dominique Sportiche, Steven Stich and Iris Trinkler for very useful discussion and comments.