Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 39 (2016) 4–14

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

What is ‘wrong’ in a neural model

Alessio Plebe

Department of Cognitive Science, University of Messina, Italy

Received 2 April 2015; received in revised form 3 September 2015; accepted 20 December 2015
Available online 14 January 2016

Abstract

Neural computation has an influential role in the study of human capacities and behaviors. It has been the dominant approach in the vision science of the last half century, and it is currently one of the fundamental methods of investigation for most higher cognitive functions. Yet, neurocomputational approaches to moral behavior are lacking. Computational modeling in general has been scarcely pursued in morality, and existent non-neural attempts have failed to account for the mental processes involved in morality. In this paper we argue that recently the situation has evolved in a way that subverted the insufficient knowledge on the basic organization of moral cognition in brain circuits, making the project of modeling morality in neurocomputational terms feasible. We will present an original architecture that combines reinforcement learning and Hebbian learning, aimed at simulating forms of moral behavior in a simple artificial context. The relationship between language and morality is controversial. In the analytic tradition of philosophy, morality is essentially the language of morals. On the other side, current cognitive ethology has shown how non human species display behaviors that are surprisingly similar to those prescribed by human ethics. Nevertheless, morality in humans is deeply entrenched with language, and the semantics of words like ‘wrong’ resists consensual explanations. The model here proposed includes an auditory processing pathway, with the purpose of showing how the coding of “wrong”, even if highly simplified with respect to its rich content in natural language, can emerge in the course of moral learning.

Keywords: Moral cognition; Neural computation; Orbitofrontal cortex; Amygdala; Self-organization

1. Introduction

Neural computation has an extraordinarily influential role in the study of several human capacities and behaviors, however the field of neurocomputational models of morality is almost unexplored yet, a failure mostly due to the lack of empirical brain information.

On the other hand, there have been computational approaches oriented toward an understanding of morality different from neurocomputation, we will briefly review two main directions: formal logic and the so-called Universal Moral Grammar (Mikhail, 2009). It will be shown that both lines of research, despite their merits, will fail in giving an account of the mental processes involved during moral cognition.

In this paper we argue that in the past decade the situation has evolved in a way that makes the project of modeling morality in neurocomputational terms feasible. Recent developments in simulating emotional responses and decision making are already offering important frameworks that we daresay able to support the project of modeling morality. The existing models deemed closer to what pertains to morality will be shortly reviewed.

By combining neural circuits for emotional responses and decision making, with simulated cortical sensorial areas, a first simple moral model has been developed (Plebe, 2014). It is the basis of the model here presented. The main extension with respect to the previous model is the inclusion of the auditory pathway, in order to explore the emergence of the lexical meaning of moral terms. Moral

http://dx.doi.org/10.1016/j.cogsys.2015.12.012
1389-0417/

behavior in humans is unquestionably related with the language of morals and value, with "wrong" the word that mostly conveys negative sentiments against actions morally sanctioned. In the neural model the meaning of "wrong" will emerge with reference to the action of stealing, the only possible moral violation in its simple artificial world.

The valuable advantage of modeling morality in a neurocomputational framework is the possibility of addressing specifically the brain components which recently have been identified as the basic support for moral cognition, in particular the areas involved in emotional driven decision making.

2. Other approaches to moral computing

Two computational accounts of morality, different from neurocomputation, will be briefly reviewed here. Both assume a strong link between morality and language, even if under different perspectives. It should be added that hard and fast claims about the necessity of language for moral behavior are debatable. Cognitive ethology has shown how other species display behaviors that are surprisingly similar to those prescribed by human ethics (Bekoff & Pierce, 2009), like compassion (Douglas-Hamilton, Bhalla, Wittemyer, & Vollrath, 2006), awareness of harm and fairness (Bekoff, 2001). Our position is that the complex structure of morality in human is language dependent indeed, however, at the core of moral cognition there are neural mechanisms shared by other animals, perhaps mammals only, which are overlooked in the two computational accounts here reviewed.

The first, with the longest tradition, has been aimed at including morality within formal logic. Hare (1952) assumed moral sentences to belong to the general class of prescriptive languages, for which meaning come in two components: the *phrastic* which captures the state to be the case, or command to be made the case, and the *neustic* part, that determines the way the sentence is nodded by the speaker. While Hare did not provide technical details of his idea for prescriptive languages, in the same years Von Wright (1951) developed deontic logic, the logical study of normative concepts in language, with the introduction of the monadic operators $O(\cdot)$, $F(\cdot)$, and $P(\cdot)$ for expressing obligation, prohibition and permission. It is well known that all the many attempts in these directions engender a set of logical and semantic problems, the most severe is the Frege–Geach embedding problem (Geach, 1965). Since the semantics of moral sentences is determined by a non-truth-apt component, like Hare's neustic, it is unclear how they can be embedded into more complex propositions, for example conditionals. This issue is related with the exclusion of mental processes within the logic formalism, and in fact viable solutions are provided by proponents of expressivism, the theory that moral judgments express attitudes of approval or disapproval, attitudes that pertain to the mental world.

One of the best available attempts in this direction has been given by Blackburn (1988) with variants of the deontic operators, like $H!(\cdot)$ and $B!(\cdot)$, that merely express attitudes regards their argument: "Hooray!" or "Boo!". Every expressive operator has its descriptive equivalent, given formally by the $|\cdot|$ operation. An alternative has been proposed by Gibbard (1990) as an extension to possible worlds semantics, defining an equivalent expressivist friendly concept, that of *factual-normative worlds* $\langle W, N\rangle$. Inside $\langle W, N\rangle$ W is an ordinary Kripke–Stalnaker possible world, while N, the system of norms, is characterized by a family of predicates like N-forbidden, N-required. If a moral sentence S is N-permitted in $\langle W, N\rangle$ then it is said to hold in that factual-normative world. Both proponents acknowledge the need of moving toward a mental inquire, but their aim did never translated into an effective attempt to embed genuine mental processes in a logic system.

The second account here sketched, was apparently motivated by filling the gap left by formal logic, the lack of the mental processes in morality. The idea that there exists a Universal Moral Grammar, that rules human moral judgments in analogy with Chomsky's Universal Grammar, was proposed several decades ago (Rawls, 1971), but has been disregarded until recently, when resuscitated by Mikhail (2009), who fleshed it out in great detail.

His fragment of Universal Moral Grammar is entirely fit to the "trolley dilemma", the famous mental experiment invented by Foot (1967), involving the so-called doctrine of the double effect, which differentiates between harm caused as means and harm caused as a side effect, like deviating a trolley killing one person but saving more lives. Mikhail refined importantly the trolley dilemma, by inventing twelve subcases that catch subtle differences. The model he developed had the purpose of computing the same average responses given by subjects on the twelve trolley subcases. It is conceived in broad analogy with a grammatical parser, taking as input a structured description of the situation and a potential action, the moral grammar, and producing as output the decision if the potential action is permissible, forbidden, or obligatory. At the core of the grammar there is a "moral calculus", including rewriting rules from actions to moral effects.

The rules are carefully defined in compliance with American jurisprudence, therefore this grammatical approach looks like a potential alternative to logical models of jurisprudence, but it is claimed to simulate the mental processes of morality. Unfortunately nothing in his model is able to support such claim. The incoherence is that the focus in the development of Mikhail is in the descriptive adequacy, the simplicity, and the formal elegance of the model, without any care on the mental plausibility. This is correct for an external epistemology, which was probably the original position of Rawls (Mallon, 2008). But a model constructed on a strict external project, and in analogy with a well established mathematical framework (formal grammar) could well have principles quite at odds with anything that is subserved by a specific mental mechanism.