# A Neural Model of Moral Decisions

Alessio Plebe
*Department of Cognitive Science, University of Messina, Messina, Italy*

Keywords: Cortical Model, Moral Cognition, Emotion Modeling, Decision Making.

Abstract: In this paper a neural model of moral decisions is proposed. It is based on the fact, supported by neuroimaging studies as well as theoretical analysis, that moral behavior is supported by brain circuits engaged more generally in emotional responses and in decision making. The model has two components, the first is composed by artificial counterpart of the orbitofrontal cortex, connected with sensorial cortical sheets and with the ventral striatum, the second by the ventromedial prefrontal cortex, that evaluate representations of values from the orbitofrontal cortex, comparing with negative values, encoded in the amygdala. The model is embedded in a simple environmental context, in which it learns that certain actions, although potentially rewarding, are morally forbidden.

## 1 INTRODUCTION

Despite the extraordinarily influential role of neural computation in the investigation of many human behaviors and capacities, no neural model for morality has yet been developed. It is not surprising, since until recently the coverage of empirical brain information about moral cognition was scarce and patchy. Since (Greene et al., 2001) directed neuroimaging studies explicitly to moral cognition the situation has significantly improved, and the current knowledge, although far from complete, is sufficient for starting a project of moral modeling. It is the purpose of this work. It will start to fill a gap inside the current trend in the study of human morality, where traditional philosophical speculation has been supplemented by a plurality of perspectives: from psychology, economics, neuroscience, anthropology, sociology. Neural computation was still missing.

This unparalleled shift in the study of morality has been described by more than one philosopher as the "empirical turn" (Nichols, 2004; Doris and Stich, 2005; Prinz, 2008), and several scholars are fostering an even more radical approach to the science of moral behavior, rooted in the understanding of the relevant brain mechanisms (Verplaetse et al., 2009; Churchland, 2011).

Two of the most important realizations to emerge from all the empirical studied done so far, are that there is no unique moral module, and that relatively consistent set of brain areas that become engaged during moral reasoning are also related to emotions, and decision making (Greene and Haidt, 2002; Moll et al., 2005; Casebeer and Churchland, 2003).

Decisions are continuously faced by the brain in everyday life, from simple motor control up to long term planning, and few of them specifically involve moral judgments. Even between actions that we may judge as "wrong" or "good", to establish a clear cut between moral norms and social conventions is not a simple and straightforward task (Kelly et al., 2007). The theoretical view embraced by this model is neo-sentimentalism. It is a view within a philosophical tradition that goes back to (Hume, 1740), which relates moral properties to certain emotions in an essential way, and construes morality as a set of prescriptive sentiments, where sentiment denotes the disposition of the subject to the relevant emotion (Nichols, 2004; Prinz, 2008).

For this reason the model here proposed is based on circuits that encode emotions, and that performs decisions on value-based representations. While neurocomputational approaches to morality are still lacking, there are indeed a number of existing models that verge on emotions and decision making, which have been a guiding reference for the development here presented. The GAGE model (Wagar and Thagard, 2004) assembles groups of artificial neurons corresponding to the ventromedial prefrontal cortex, the hippocampus, the amygdala, and the nucleus accumbens, in implementing the somatic-marker effect: encoding of feelings that have become associated

Plebe A..
A Neural Model of Moral Decisions.
DOI: 10.5220/0005032001110118
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2014), pages 111-118
ISBN: 978-989-758-054-3

through experience with the predicted long-term outcomes of certain responses (Damasio, 1994). In the ANDREA model (Litt et al., 2008) the orbitofrontal cortex, the dorsolateral prefrontal cortex, and the anterior cingulate cortex interact with basal ganglia and the amygdala in reproducing the human hypersensitivity to losses over equivalent gains (Kahneman and Tversky, 1979). The overall architecture of these models shares similarities with those of (Frank and Claus, 2006; Frank et al., 2007), in which the orbitofrontal cortex interacts with the basal ganglia to produce dichotomic on/off decisions.

The proposed model is made of several simulated cortical and subcortical areas, described in detail in §2. It is embedded in a simplified world, which can be experienced through vision and taste. There are two possible kinds of objects in the scene, only one is edible, like a fruit. Collecting fruits is not allowed everywhere, there are areas where it is forbidden, and any violation will call into action an angry face, visible in the scene. Results about the ability of the model to learn this simple moral rule, and act accordingly, will be shown in §3.

## 2 DESCRIPTION OF THE MODEL

The overall model is shown in Fig. 1. It is composed by a series of sheets with artificial neural units, labeled with the acronym of the brain structure that is supposed to reproduce. It is implemented using the *Topographica* neural simulator (Bednar, 2009), and each cortical sheet adheres to the LISSOM (*Laterally Interconnected Synergetically Self-Organizing Map*) concept (Sirosh and Miikkulainen, 1997). In a LISSOM sheet of neurons, the activation of each neuron is due to the combination of afferents and excitatory and inhibitory lateral connections, as detailed below.

There are two main circuits that learn the emotional component that contributes to the evaluation of potential actions. A first one comprises the orbitofrontal cortex, with its processing of sensorial information, reinforced with positive perspective values by the loop with the ventral striatum. The second one shares the representations of values from the orbitofrontal cortex, which are evaluated by the ventromedial prefrontal cortex against conflicting negative values, encoded by the closed loop with the amygdala. The subcortical sensorial components comprise LGN at the time when seeing the main scene, the LGN deferred in time, when a possibly angry face will appear, and the taste information.

### 2.1 Equations at the Single Neuron Level

The basic equation of the LISSOM describes the activation level $x_i$ of a neuron $i$ at a certain time step $k$:

$$x_i^{(k)} = f\left(\gamma_{\mathrm{A}}\vec{a}_i \cdot \vec{v}_i + \gamma_{\mathrm{E}}\vec{e}_i \cdot \vec{x}_i^{(k-1)} - \gamma_{\mathrm{H}}\vec{h}_i \cdot \vec{x}_i^{(k-1)}\right) \tag{1}$$

The vector fields $\vec{v}_i$, $\vec{e}_i$, $\vec{x}_i$ are circular areas of radius $r_{\mathrm{A}}$ for afferents, $r_{\mathrm{E}}$ for excitatory connections, $r_{\mathrm{H}}$ for inhibitory connections. The vector $\vec{a}_i$ is the receptive field of the unit $i$. Vectors $\vec{e}_i$ and $\vec{h}_i$ are composed by all connection strengths of the excitatory or inhibitory neurons projecting to $i$. The scalars $\gamma_{\mathrm{A}}$, $\gamma_{\mathrm{E}}$, $\gamma_{\mathrm{H}}$, are constants modulating the contribution of afferents, excitatory, inhibitory and backward projections. The function $f$ is a piecewise linear approximation of the sigmoid function, $k$ is the time step in the recursive procedure. The final activation of neurons in a sheet is achieved after a small number of time step iterations, typically 10.

All connection strengths adapt according to the general Hebbian principle, and include a normalization mechanism that counterbalances the overall increase of connections of the pure Hebbian rule. The equations are the following:

$$\Delta\mathbf{a}_{r_{\mathrm{A}},i} = \frac{\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}}{\|\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}\|} - \mathbf{a}_{r_{\mathrm{A}},i}, \tag{2}$$

$$\Delta\mathbf{e}_{r_{\mathrm{E}},i} = \frac{\mathbf{e}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}}{\|\mathbf{a}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}\|} - \mathbf{e}_{r_{\mathrm{E}},i}, \tag{3}$$

$$\Delta\mathbf{i}_{r_{\mathrm{I}},i} = \frac{\mathbf{i}_{r_{\mathrm{I}},i} + \eta_{\mathrm{I}} x_i \mathbf{x}_{r_{\mathrm{I}},i}}{\|\mathbf{i}_{r_{\mathrm{I}},i} + \eta_{\mathrm{I}} x_i \mathbf{x}_{r_{\mathrm{I}},i}\|} - \mathbf{i}_{r_{\mathrm{I}},i}, \tag{4}$$

where $\eta_{\{\mathrm{A,E,I}\}}$ are the learning rates for the afferent, excitatory, and inhibitory weights, and $\|\cdot\|$ is the $\mathrm{L}^1$-norm.

### 2.2 Orbitofrontal Circuit

The first circuit in the model learns the positive reward in eating fruits. The orbitofrontal cortex is the site of several high level functions (Rolls, 2004), in this model information from the visual stream and taste have been used. There are neurons in the orbitofrontal cortex that respond differentially to visual objects depending on their taste reward (Rolls et al., 1996), and others which respond to facial expressions (Rolls et al., 2006), involved in social decision making (Damasio, 1994; Bechara et al., 1994). For (Prehn and Heekeren, 2009) the role of the orbitofrontal cortex in moral judgment is the representation of the expected value of possible outcomes of a behavior in regards to rewards and punishments.

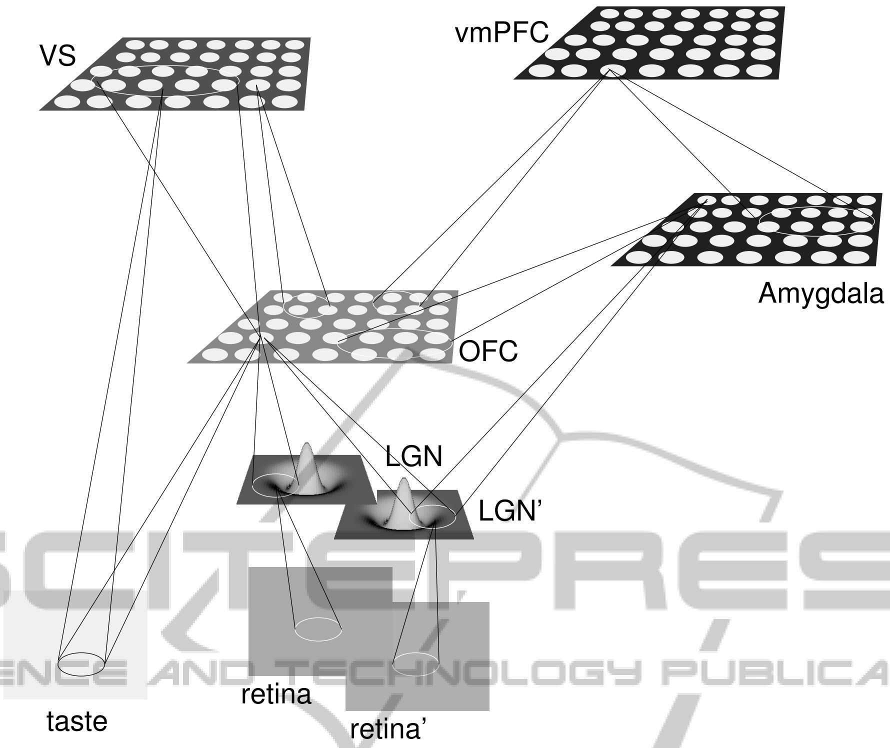

Figure 1: Overall scheme of the model, composed by LGN (*Lateral Geniculate Nucleus*), V1 (*Primary Visual Area*), OFC (*OrbitoFrontal Cortex*), VS (*Ventral Striatum*), Amyg (*Amygdala*), vmPFC (*ventromedial PreFrontal Cortex*).

In all the subsequent equations the superscript of time step s$k$ will be omitted, for sake of readability, and substituted by the name of the sheet, using the abbreviations $L$ and $L'$ for, respectively, the output of the LGN at the time when seeing the main scene, and the output of the LGN deferred in time, and the abbreviation $T$ for the taste signal.

The equation of the activation of a neural unit in the OFC layer is the following:

$$\begin{aligned} x^{(\mathrm{OFC})} = f\Big(&\gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})}\vec{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})}\cdot\vec{v}_{r_{\mathrm{A}}}^{(\mathrm{V1})}+\\ &\gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\mathrm{L}')}\vec{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\mathrm{L}')}\cdot\vec{v}_{r_{\mathrm{A}}}^{(\mathrm{L}')}+\\ &\gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\mathrm{T})}\vec{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\mathrm{T})}\cdot\vec{v}_{r_{\mathrm{A}}}^{(\mathrm{T})}+\\ &\gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{VS})}\vec{b}_{r_{\mathrm{B}}}^{(\mathrm{OFC})}\cdot\vec{v}_{r_{\mathrm{B}}}^{(\mathrm{VS})}+\\ &\gamma_{\mathrm{E}}^{(\mathrm{OFC})}\vec{e}_{r_{\mathrm{E}}}^{(\mathrm{OFC})}\cdot\vec{x}_{r_{\mathrm{E}}}^{(\mathrm{OFC})}-\\ &\gamma_{\mathrm{H}}^{(\mathrm{OFC})}\vec{h}_{r_{\mathrm{H}}}^{(\mathrm{OFC})}\cdot\vec{x}_{r_{\mathrm{H}}}^{(\mathrm{OFC})}\Big)\end{aligned} \qquad (5)$$

which is a specialization of the general equation (1). Here, for better readability, the unit index $i$ and time step $k$ have been omitted. There are three sensorial afferents: $\vec{v}_{r_{\mathrm{A}}}^{(\mathrm{V1})}$ from the visual cortex V1, $\vec{v}_{r_{\mathrm{A}}}^{(\mathrm{L})}$ from the lateral geniculate nucleus of the thalamus, and the taste sensorial input $\vec{v}_{r_{\mathrm{A}}}^{(\mathrm{T})}$, each in a sensorial area $r_{\mathrm{A}}$ corresponding to the receptive field of the unit in OFC. The fourth afferent, $\vec{v}_{r_{\mathrm{B}}}^{(\mathrm{VS})}$, is the backprojection from the VS loop that will be described next. The visual pathway is simplified in a single area, V1, with the following equation:

$$\begin{aligned} x^{(\mathrm{V1})} = h\Big(&\gamma_{\mathrm{A}}^{(\mathrm{V1}\leftarrow\mathrm{L})}\vec{a}_{r_{\mathrm{A}}}^{(\mathrm{V1}\leftarrow\mathrm{L})}\cdot\vec{v}_{r_{\mathrm{A}}}^{(\mathrm{L})}+\gamma_{\mathrm{E}}^{(\mathrm{V1})}\vec{e}_{r_{\mathrm{E}}}^{(\mathrm{V1})}\cdot\vec{x}_{r_{\mathrm{E}}}^{(\mathrm{V1})}-\\ &\gamma_{\mathrm{H}}^{(\mathrm{V1})}\vec{h}_{r_{\mathrm{H}}}^{(\mathrm{V1})}\cdot\vec{x}_{r_{\mathrm{H}}}^{(\mathrm{V1})}\Big)\end{aligned} \qquad (6)$$

which differs from equation (1) in that the nonlinear function $h$ has an adaptive threshold $\theta$, dependent on the average activity of the unit, using:

$$\theta^{(k)} = \theta^{(k)} + \lambda\left(\bar{x}^{(\mathrm{V1})} - \mu\right) \qquad (7)$$

where $\bar{x}^{(\mathrm{V1})}$ is a smoothed exponential average in time of the activity, and $\lambda$ and $\mu$ fixed parameters. This feature simulates the biological adaptation that allows the development of stable topographic maps organized by preferred retinal location and orientation (Stevens et al., 2013). The output of LGN is given by:

$$x^{(L)} = f\left(\frac{\gamma_O\left(\vec{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{N}})} - \vec{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{W}})}\right)\cdot\vec{v}_{r,c}}{\beta+\gamma_S\vec{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{S}})}\cdot\vec{x}_{\mathrm{S}}^{(L)}}\right) \qquad (8)$$

approximating the combined contribution of ganglion cells and LGN with a positive center and negative surround, by differences of two Gaussian $\vec{g}^{(\sigma_\mathrm{N})}$ and $\vec{g}^{(\sigma_\mathrm{W})}$, with the denominator term acting as contrast-gain control (Stevens et al., 2013). The bidimensional coordinates $r$ and $c$ refers to the retinal photoreceptors, and $\vec{x}_\mathrm{S}^{(L)}$ are the suppressive connection field of the given unit. It holds $\sigma_\mathrm{N} < \sigma_\mathrm{S} < \sigma_\mathrm{W}$.

OFC has forward and feedback connections with the Ventral Striatum, VS, which is the crucial center for various aspects of reward processes and motivation (Haber, 2011). VS in the model is a crude simplification of this complex area, and does not reproduce the details of its direct and reciprocal connection with the dopaminergic neurons centers. It is implemented by the following equation:

$$x^{(\mathrm{VS})} = f\Bigg(\gamma_\mathrm{A}^{(\mathrm{VS}\leftarrow\mathrm{OFC})}\vec{a}_{r_\mathrm{A}}^{(\mathrm{VS}\leftarrow\mathrm{OFC})}\cdot\vec{v}_{r_\mathrm{A}}^{(\mathrm{OFC})} + \gamma_\mathrm{A}^{(\mathrm{VS}\leftarrow\mathrm{T})}\vec{a}_{r_\mathrm{A}}^{(\mathrm{VS}\leftarrow\mathrm{T})}\cdot\vec{v}_{r_\mathrm{A}}^{(\mathrm{T})} + \gamma_\mathrm{E}^{(\mathrm{VS})}\vec{e}_{r_\mathrm{E}}^{(\mathrm{VS})}\cdot\vec{x}_{r_\mathrm{E}}^{(\mathrm{VS})} - \gamma_\mathrm{H}^{(\mathrm{VS})}\vec{h}_{r_\mathrm{H}}^{(\mathrm{VS})}\cdot\vec{x}_{r_\mathrm{H}}^{(\mathrm{VS})}\Bigg) \tag{9}$$

The afferent signals $\vec{v}^{(\mathrm{OFC})}$ come from equation (5), $\vec{v}^{(\mathrm{T})}$ is the taste signal. The output $x^{(\mathrm{VS})}$ computed in (9) will close the loop into the prefrontal cortex with equation (5).

## 2.3 Ventromedial Circuit

The second main circuit in the model is based on the ventromedial prefrontal cortex, vmPFC, and its connections from OFC and the amygdala. The ventromedial prefrontal cortex is long since known to play a crucial role in emotion regulation and social decision making (Bechara et al., 1994; Damasio, 1994). More recently it has been proposed that the vmPFC may encode a kind of common currency enabling consistent value based choices between actions and goods of various types (Gläscher et al., 2009). It is involved in the development of morality, in a study (Decety et al., 2012) older participants showed significant stronger coactivation between vmPFC and amygdala when attending to scenarios with intentional harm, compared to younger subjects. The amygdala is the primary mediator of negative emotions, and responsible for learning associations that signal a situation as fearful (LeDoux, 2000). In the model it is used specifically for capturing the negative emotion when seeing the angry face, a function well documented in the amygdala (Boll et al., 2011).

vmPFC is implemented in MONE using the standard equation (1), as follows:

$$x^{(\mathrm{vFC})} = f\Bigg(\gamma_\mathrm{A}^{(\mathrm{vFC}\leftarrow\mathrm{OFC})}\vec{a}_{r_\mathrm{A}}^{(\mathrm{vFC}\leftarrow\mathrm{OFC})}\cdot\vec{v}_{r_\mathrm{A}}^{(\mathrm{OFC})} + \gamma_\mathrm{A}^{(\mathrm{vFC}\leftarrow\mathrm{Amy})}\vec{a}_{r_\mathrm{A}}^{(\mathrm{vFC}\leftarrow\mathrm{Amy})}\cdot\vec{v}_{r_\mathrm{A}}^{(\mathrm{Amy})} + \gamma_\mathrm{E}^{(\mathrm{vFC})}\vec{e}_{r_\mathrm{E}}^{(\mathrm{vFC})}\cdot\vec{x}_{r_\mathrm{E}}^{(\mathrm{vFC})} - \gamma_\mathrm{H}^{(\mathrm{vFC})}\vec{h}_{r_\mathrm{H}}^{(\mathrm{vFC})}\cdot\vec{x}_{r_\mathrm{H}}^{(\mathrm{vFC})}\Bigg) \tag{10}$$

The afferent signals $\vec{v}^{(\mathrm{OFC})}$ come from equation (5), while $\vec{v}^{(\mathrm{Amy})}$, the connection from Amygdala, is given from the following equation:

$$x^{(\mathrm{Amy})} = f\Bigg(\gamma_\mathrm{A}^{(\mathrm{Amy}\leftarrow\mathrm{OFC})}\vec{a}_{r_\mathrm{A}}^{(\mathrm{Amy}\leftarrow\mathrm{OFC})}\cdot\vec{v}_{r_\mathrm{A}}^{(\mathrm{OFC})} + \gamma_\mathrm{A}^{(\mathrm{Amy}\leftarrow\mathrm{L}')}\vec{a}_{r_\mathrm{A}}^{(\mathrm{Amy}\leftarrow\mathrm{L}')}\cdot\vec{v}_{r_\mathrm{A}}^{(\mathrm{L}')} + \gamma_\mathrm{E}^{(\mathrm{Amy})}\vec{e}_{r_\mathrm{E}}^{(\mathrm{Amy})}\cdot\vec{x}_{r_\mathrm{E}}^{(\mathrm{Amy})} - \gamma_\mathrm{H}^{(\mathrm{Amy})}\vec{h}_{r_\mathrm{H}}^{(\mathrm{Amy})}\cdot\vec{x}_{r_\mathrm{H}}^{(\mathrm{Amy})}\Bigg) \tag{11}$$

The afferent signals $\vec{v}^{(\mathrm{OFC})}$ come from equation (5), while $\vec{v}^{(\mathrm{L}')}$ is a direct reading of face from the visual afferents in the thalamus, delayed in time with respect to the ordinary visual scene. The activation given from equation (11) will loop inside the vmPFC by equation (10).

## 2.4 Decisions in the Ventromedial Area

A method of analysis has been carried out for the identification of decisions as population coding of neural activation in the vmPFC map. Let us introduce the following function:

$$x_i(e) : E \in \mathcal{E} \to \mathbb{R}; \quad s \in E \in \mathcal{E}, \tag{12}$$

that gives the activation $x$ of a generic neuron $i$ in vmPFC in response to an environmental condition $e$. This condition is an instance of a class $E$, belonging to the set of all classes of conditions $\mathcal{E}$. In this experiment $\mathcal{E} = \{E_1, E_2, E_3\}$, where $E_1$ is the set of situations where an eatable object is freely available in the scene, in $E_2$ a fruit is still in the scene, but forbidden, in $E_3$ there is a neutral object in the scene. For a class $E \in \mathcal{E}$ we can define the two sets:

$$X_{E,i} = \big\{x_i(e_j) : e_j \in E\big\}; \tag{13}$$

$$\overline{X}_{E,i} = \big\{x_i(e_j) : e_j \in E' \neq E \in \mathcal{E}\big\}. \tag{14}$$

We can then associate to the class $E$ a set of neurons in the map, by ranking it with the following function:

$$r(E,i) = \frac{\mu_{X_{E,i}} - \mu_{\overline{X_{E,i}}}}{\sqrt{\frac{\sigma_{X_{E,i}}}{|X_{E,i}|} + \frac{\sigma_{\overline{X_{E,i}}}}{|\overline{X}_{E,i}|}}}, \tag{15}$$

Table 1: Main parameters of all model maps. Radius values are normalized in range $[0\dots2]$.

| layer | size | $r_{\mathrm{A}}$ | $r_{\mathrm{A'}}$ | $r_{\mathrm{E}}$ | $r_{\mathrm{H}}$ | $\gamma_{\mathrm{A}}$ | $\gamma_{\mathrm{A'}}$ | $\gamma_{\mathrm{E}}$ | $\gamma_{\mathrm{H}}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| LGN | $24\times24$ | 0.3 | - | - | - | - | - | - | - | - |
| V1 | $22\times22$ | 0.1 | - | 0.8 | 0.4 | 2.0 | - | 1.0 | 0.3 | |
| OFC | $16\times16$ | 0.4 | 0.1 | 0.1 | 0.5 | 0.8 | 0.4 | 1.4 | 1.6 | |
| VS | $8\times8$ | 0.5 | 0.3 | 0.6 | 0.3 | 0.6 | 0.4 | 1.4 | 0.4 | |
| Amyg | $8\times8$ | 0.6 | 0.5 | 0.5 | 0.6 | 0.6 | 0.3 | 0.5 | 1.5 | |
| vmPFC | $12\times12$ | 0.4 | 0.5 | 0.4 | 0.2 | 2.0 | $-0.5$ | 1.2 | 1.7 | |

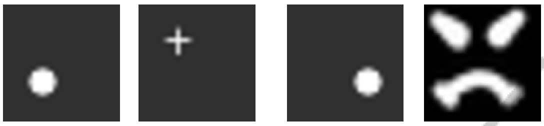

Figure 2: The visual inputs of the model. From the left to the right: an edible object, possibly an apple, a + shaped neutral object, the edible object in the forbidden area (the bottom right quadrant of the scene), which is followed by the sad and angry schematic face.

where $\mu$ is the average and $\sigma$ the standard deviation of the values in the two sets, and $|\cdot|$ is the cardinality of a set. Now the following relation can be established as the population code of a condition class $E$:

$$\begin{aligned} p(E) : \mathcal{E} \to \{\langle i_1, i_2, \cdots, i_M\rangle : \\ r(E,i_1) > r(E,i_2) > \cdots > r(E,i_M)\}, \end{aligned} \quad (16)$$

where $M$ is a given constant, typically one order of magnitude smaller than the number of neurons in the map The population code $p(E)$ computed with (16) is used to take a decision $d \in \mathcal{D}$ according to the status $e$ of the environment. In this experiment $\mathcal{D} = \{d_1, d_2\}$, where $d_1$ is the decision to collect and eat the object, $d_2$ is the decision to ignore it.

$$d(e) = m\left(\arg\max_{E\in\mathcal{E}}\left\{\sum_{j=1\cdots E}\alpha^j x_{p(E)_j}(s)\right\}\right), \quad (17)$$

where $p(E)_j$ denotes the $j$-th element in the ordered set $p(E)$; $\alpha$ is a constant that is close, but smaller, than one; $m(\cdot)$ is a mapping function from environmental categories to decision:

$$\begin{aligned} E_1 &\to d_1; \\ E_2 &\to d_2; \\ E_3 &\to d_2. \end{aligned} \quad (18)$$

# 3 RESULTS AND DISCUSSION

The artificial moral brain architecture just described is exposed to a series of situations that simulate highly simplified contexts, and the appropriate action is gradually learned. Some actions are charged with important survival reward, but in some cases may cause detriment to others. Their angry reaction will lead to learn that that action is "wrong".

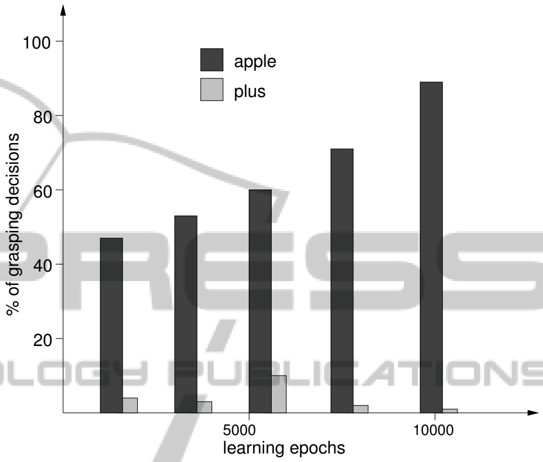

Figure 3: Percentage of grasping actions selected by the vmPFC model map, for the apple and the + shaped neutral object, at different epochs of the development.

The main input to the model is a visual scene, examples are shown in the Fig. 2. Our artificial subject is unfamiliar with the objects, she can realize how pleasant fruits are to eat, thanks to its taste perception. This sensorial input is simply a matrix $2\times2$, in which the ratio of the upper row to the lower row signal how pleasant the taste is. Fruits in the bottom right quadrant may belong to a member of the social group, and to collect these fruits would be a violation of her/his property, that would trigger an immediate reaction of sadness and anger. This reaction is perceived in the form of a face with a marked emotion, as the one in the rightmost position in Fig. 2

## 3.1 First Learning Stage

This phases of development includes an early stage of formation of V1, the visual system, with elongated patterns as inputs, followed by item the good food recognition stage, in which the stimuli are the two types of objects, in all possible positions, and their taste.

Learning is always ruled by equations (2), (3),

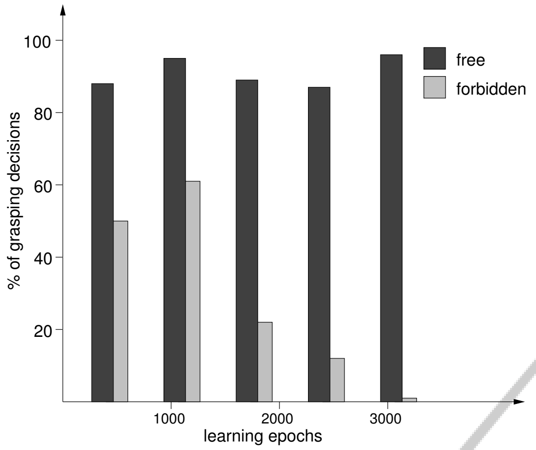

Figure 4: Percentage of grasping actions selected by the vmPFC model map for the apple, placed in the free, or om the forbidden places, ad different epochs of the development.

and (4). applied to the relevant connections. The development of V1, involving equation (6) only, setup the main systems of organization in the primary visual cortex, with arrangement of orientation tuned neurons, similar to that described in (Plebe and Domenella, 2007).

Then the OFC, VS, and vmPFC areas of the model become plastic, and learn their connections of equations (5), (9), and (10). This set of equations is an implicit reinforcement learning, where the reward is not imposed externally, but acquired by the OFC map, through its taste sensorial input. The amygdala has no interaction during these stages.

The coding in vmPFC model map is the decision made to grasp or not to grasp the object, the percentage of decision to grasp, at various learning steps, is shown in Fig. 3. When the object is an apple, grasping gradually become the prevailing choice, that reaches to 60% after 5000 learning epochs, and to 90% at the end of this learning phase. Occurrence of grasping is instead low for the non rewarding object, and become meaningless, below 5%, at the end of the learning phase.

## 3.2 Moral Learning

In this second phase the model receives additional experiences, that of the moral emotion learning, with the objects as stimuli, followed by an image in which there could be the angry face. This face will pop up only when an object of the first kind, the apple, appears in the right bottom quadrant in the scene. This is a sort of private property, and the owner reacts with sadness and anger when his fruit has been grasped.

Now the amygdala gets inputs from both the OFC map and directly from the thalamus, when the angry face appears, as from equation (11), and learns its connections. In this case, there is an implicit reinforcement learning as well, with the negative reward embedded in the input projections to the amygdala.

In Fig. 4 there are the percentages of decisions to grasp the apple fruit, decoded as before from the vmPFC map. In this case, the samples of the edible object have been divided in two groups, depending on the position in the scene. It can be seen how the model develops a strong inhibition to grasp the edible objects when placed in the forbidden sector. At the end of this development phase the cases of transgression have dropped below 1%. The percentage of decisions to collect fruits inside the free area of the scene are always high in all this phase, with values above 90% at the end of the development. It can be claimed that the model has learned a moral rule, as an imperative inhibition to perform certain actions.

## 3.3 Conclusions

We have described a first attempt to simulate moral cognition in a neurocomputational model. It has significant limitations, and we think its contribution to the progress of moral science will be modest. First, the model is able to simulate only one kind of moral situation, the temptation of stealing food, and the potential consequent feelings of guilt. Since morality is a collection of several, partially dissociated mechanisms, a model must necessarily, at least in its first implementation, choose a specific one to target. Second, even in the single case of stealing, and consequent guilt, the model is missing many brain areas that are potentially involved, like the cingulate cortex and the hippocampus, to name few.

Both the design of the moral situation, and the architecture of the brain areas, derive from a compromise between manageability of the model, and the level of knowledge of the functions in brain areas potentially involved. The food stealing situation offers the advantage of adopting external signals, visual and of taste, with a well established connections in crucial areas included in the model, like the orbitofrontal cortex and the amygdala.

While the schematic external world of the model is a pale resemblance of a typical real situation of human moral decision, it is a major advance with respect to any existing neural model of decisions. For example, in the ANDREA model (Litt et al., 2008) there is only a single input that signals a gain when positive and a loss when negative; in the models of Frank and co-workers the input is a combination of 4 possible abstract cues (Frank and Claus, 2006). In this model the brain circuits for decisions and emotions have been complemented by a visual system and a simplified taste input, allowing the simulation of a schematic worlds where moral relevant events take place.

Therefore, even in its crudely simplified form, the model simulates a typical moral situation, using the relevant stimuli, and plausible neural mechanisms, in a hierarchy of areas that capture the essence of the moral decision to be done. We believe that the neurocomputational approach is an additional important path in pursuing a better understanding of morals, and this model, despite the limitations here discussed, is a valid starting point.

## REFERENCES

Bechara, A., Damasio, A. R., Damasio, H. R., and Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50:7–15.

Bednar, J. A. (2009). Topographica: Building and analyzing map-level simulations from Python, C/C++, MATLAB, NEST, or NEURON components. *Frontiers in Neuroinformatics*, 3:8.

Boll, S., Gamer, M., Kalisch, R., and Büchel, C. (2011). Processing of facial expressions and their significance for the observer in subregions of the human amygdala. *NeuroImage*, 56:299–306.

Casebeer, W. D. and Churchland, P. S. (2003). The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy*, 18:169–194.

Churchland, P. S. (2011). *Braintrust – what neuroscience tells us about morality*. Princeton University Press, Princeton (NJ).

Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. Avon Books, New York.

Decety, J., Michalska, K. J., and Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, 22:209–220.

Doris, J. M. and Stich, S. P. (2005). As a matter of fact: Empirical perspectives on ethics. In Jackson, F. and Smith, M., editors, *The Oxford Handbook of Contemporary Philosophy*. Oxford University Press, Oxford (UK).

Frank, M. J. and Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, 113:300–326.

Frank, M. J., Scheres, A., and Sherman, S. J. (2007). Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. *Philosophical transactions of the Royal Society B*, 362:1641–1654.

Gläscher, J., Hampton, A. N., and O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, 19:483–495.

Greene, J. D. and Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6:517–523.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). fMRI investigation of emotional engagement in moral judgment. *Science*, 293:2105–2108.

Haber, S. N. (2011). Neural circuits of reward and decision making: Integrative networks across corticobasal ganglia loops. In (Mars et al., 2011), pages 22–35.

Hume, D. (1740). *A Treatise of Human Nature*. Thomas Longman, London. Vol 3.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:313–327.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., and Fessler, D. M. T. (2007). Harm, affect, and the moral/conventional distinction. *Minds and Language*, 22:117–131.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, 23:155–184.

Litt, A., Eliasmith, C., and Thagard, P. (2008). Neural affective decision theory: Choices, brains, and emotions. *Cognitive Systems Research*, 9:252–273.

Mars, R. B., Sallet, J., Rushworth, M. F. S., and Yeung, N., editors (2011). *Neural Basis of Motivational and Cognitive Control*. MIT Press, Cambridge (MA).

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, 6:799–809.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press, Oxford (UK).

Plebe, A. and Domenella, R. G. (2007). Object recognition by artificial cortical maps. *Neural Networks*, 20:763–780.

Prehn, K. and Heekeren, H. R. (2009). Moral judgment and the brain: A functional approach to the question of emotion and cognition in moral judgment integrating psychology, neuroscience and evolutionary biology. In (Verplaetse et al., 2009).

Prinz, J. (2008). *The Emotional Construction of Morals*. Oxford University Press, Oxford (UK).

Rolls, E. (2004). The functions of the orbitofrontal cortex. *Biological Cybernetics*, 55:11–29.

Rolls, E., Critchley, H., Browning, A. S., and Inoue, K. (2006). Face-selective and auditory neurons in the primate orbitofrontal cortex. *Experimental Brain Research*, 170:74–87.

Rolls, E., Critchley, H., Mason, R., and Wakeman, E. A. (1996). Orbitofrontal cortex neurons: Role in olfactory and visual association learning. *Journal of Neurophysiology*, 75:1970–1981.

Sirosh, J. and Miikkulainen, R. (1997). Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation*, 9:577–594.

Stevens, J.-L. R., Law, J. S., Antolik, J., and Bednar, J. A. (2013). Mechanisms for stable, robust, and adaptive development of orientation maps in the primary visual cortex. *JNS*, 33:15747–15766.

Verplaetse, J., Schrijver, J. D., Vanneste, S., and Braeckman, J., editors (2009). *The Moral Brain Essays on the Evolutionary and Neuroscientific Aspects of Morality*. Springer-Verlag, Berlin.

Wagar, B. M. and Thagard, P. (2004). Spiking Phineas Gage: A neurocomputational theory of cognitiveaffective integration in decision making. *Psychological Review*, 111:67–79.