Biol Cybern (2015) 109:685–699
DOI 10.1007/s00422-015-0669-z

ORIGINAL ARTICLE

Biological Cybernetics

# Neurocomputational model of moral behaviour

**Alessio Plebe**[1]

Received: 30 January 2014 / Accepted: 7 November 2015 / Published online: 19 November 2015


**Abstract** Our understanding of human morality has dramatically improved in the last decades, thanks to efforts carried out with scientific methods, in addition to the traditional speculative approach. Substantial contributions and relevant empirical data have come from neuroscience, psychology, genetics, comparative ethology, anthropology, and the social sciences. In this fruitful synergy, one useful approach is still missing: computational modeling. More precisely, a neurocomputational model aimed at simulating forms of moral behavior, to our knowledge, has not yet been designed. The purpose of this work is to start filling this gap, proposing MOral Neural Engine (MONE), a model that simulates the emergence of moral cognition. The neural engine in this model is assumed to be based in frontal areas, specifically the orbitofrontal and the ventromedial prefrontal cortex, and in connections to limbic areas involved in emotions and reward, such as the ventral striatum and the amygdala. Moral cognition is probably the result of a collection of several different neural processes, activated depending on the type of moral problem, each associated with a variety of emotions. This model, in its first implementation, deals with only a single moral situation: stealing someone's food, a transgression that typically elicits guilt, learned in the model from the angry facial expressions of the victim.

**Keywords** Moral behavior · Cortical model · Reinforcement learning · Decision-making · Moral emotions

✉ Alessio Plebe
aplebe@unime.it

[1] Department of Cognitive Science, Education, and Cultural Studies, v. Concezione 8, Messina, Italy

## 1 Introduction

A radical shift in the study of human morality has taken place in the last few decades, leading more than one philosopher to describe it as the "empirical turn" (Nichols 2004; Doris and Stich 2005; Prinz 2008a). In some respects, this tendency is not surprising, it can be seen as internal to the more general view in current philosophy of mind, that theorizing should be essentially a posterior inquiry informed by relevant empirical data, primarily from neuroscience, and the behavioral and social sciences. This is indeed partially true, but morality for long has had its own special reasons for shunning empirical considerations.

The famous "open-question argument" put forward by Moore (1903) allegedly dismissed any proposition that inferred moral value from natural properties, as a logical fallacy. Even if his argument was basically flawed (Frankena 1939), Moore's non-naturalism profoundly shaped moral philosophy in his century. Even more compelling is an older similar claim by Hume (1740) that no "ought" can be derived by an "is." Scientific facts are descriptive, while moral facts are prescriptive, and since it is impossible to deduce a statement that has obligatory force from statements that are purely descriptive, moral theories cannot be pursued empirically. A cautionary affirmation distinctively peremptory, for having being devised by the champion of empiricism. Therefore, it is not surprising that other forms of human cognition, like perception, reasoning, and language, have gained a respectable place inside the empirical sciences long before morality.

Psychology has traditionally been the domain most profusive of experimental observations relevant to morals. However, throughout the twentieth century, the psychological literature on moral judgment has been largely ignored by philosophers. This was in part due to the closure to empirical interactions, for the reasons explained above, in

part because much of the psychological work seemed not to address the core of morality, rather, through Kohlberg and Piaget, attempted to delineate variations in moral cognition, like developmental changes, or gender differences. Only in the last decades has moral psychology produced research of great theoretical impact, mainly thanks to neurocognitive methodologies, as will be discussed shortly.

In the meantime, it is worth mentioning other sources of empirical observation concerning morals. One of the most important contributions comes from anthropological studies. Since mainstream moral philosophy focuses on western ethical traditions, their body of work has been highly informative, detailing the radically divergent moral outlooks found in cultures around the world (see, for example Chagnon 1974; Turnbull 1973). While one of the most valuable contributions from anthropology is to reveal the assorted variety of morals, recent cognitive ethology has shown how other species display behaviors that are surprisingly similar to those prescribed by human ethics (Douglas-Hamilton et al. 2006; Bekoff 2001; Bekoff and Pierce 2009).

But neuroscience is, first and foremost, the main source of empirical data that has come to shed a better light on the nature of moral cognition. Although moral cognitive neuroscience is still in its infancy, the wealth of data collected so far already points to a sketchy picture of how the brain supports morality (Casebeer and Churchland 2003; Moll et al. 2005). The cross-field integration with psychology unclosed the door to the—now famous —trolley mental experiments (Greene et al. 2001; Greene and Haidt 2002). Churchland (2011) outlined a philosophical synthesis consistent with current neuroscientific data, integrated with other empirical sources as genetics and evolutionary biology. The neural basis of morals is at the root of MONE, and therefore it will be detailed in a specific section.

The range of empirical approaches just listed belong to the typical synergy of cognitive science. A main one is missing: computational modeling. More precisely, as far as we know, a neurocomputational model that aimed at simulating forms of moral behavior has not yet been designed. This is exactly the purpose of MOral Neural Engine (MONE),[1] the model here proposed.

## 2 Computational models related to morality

While neurocomputational approaches to morality are still lacking, there are indeed a number of computational models that are to some extent related with the study of morality, some of these have served as inspiration for the development of MONE. In briefly reviewing the state of the art, three main categories of models can be identified.

[1] With explicit reference to the imperative of the Latin verb `monere`.

### 2.1 Machine morality

There is a growing interest in expanding the scope of autonomous agents, so that they honor the broader set of values and laws humans demand of human moral agents. The motivations of this body of research is essentially pragmatic. With the spreading use of autonomous systems, there is an increasing risk of danger and inconvenience. Wallach and Allen (2008) point out that Greene's famous trolley cases can become real dilemmas for autonomous agents, with the advent of modern driverless train systems. An even more straightforward example of technological agents that need some sort of moral control, are the remotely operated vehicles being deployed militarily. Even if driven by practical needs, the engineering of morality entails interesting and challenging philosophical issues (Gunkel 2012), and is a field of research that could squarely intersect with the purposes here pursued. However, the main difference is that in machine morality there is no need to implement an architecture similar to the human brain. The only requirement is that the external behavior, within the set of actions designed in the agent, respects certain moral norms. On the contrary, MONE aims at replicating, as faithfully as possible, the structure that in the brain gives rise to moral cognition, although limited to a very narrow simulated world. It might be that the architecture designed in MONE could reveal a valid basis for developing moral engines in artificial agents, but this is out of the scope of the present work.

### 2.2 Economic models

Game theory is the main mathematical method in economics for modeling competing behaviors of social agents. Within the large collection of possible economic behaviors, there are decisions strongly affected by the moral asset of the agent. In particular, the most studied case is that where the decision is between two conflicting choices, one selfish and the other altruistic. The best prototype is the well-known Prisoner's Dilemma, the schematic testbed of hundreds of economic models. Since the seminal work of Hamilton (1963), game theory has been used in exploring the evolution of altruism, under the hypothesis of kin selection. The main outcome is the well-known Hamilton's rule, a condition to be satisfied for the altruistic train to be spread:

$$rb - c > 0 \tag{1}$$

where $b$ is the fitness benefits provided to partners, $c$ is the fitness cost of helping, and $r$ is the coefficient of genetic relatedness between helper and helped (Hamilton 1964). During the last decades, Hamilton's basic model has been refined and extended in several directions, including cases of coop-

eration and altruistic helping between non-kin individuals (Lehmann and Keller 2006).

There are several reasons why all these models are not relevant for MONE. First of all, altruism and cooperation are clearly related with morality, but intersect only a narrow aspect of moral cognition. Second, economic models lack any detail about the cognitive processes that make the decision between a selfish or altruistic choice possible. Finally, in this work, the purpose is to model the brain of a single individual only, engaged in moral cognition. We expect in future developments to extend MONE by endowing it with communicative capabilities and use it as a basis for a population of interacting moral agents.

### 2.3 Neurocomputational models of decisions and emotions

Not every human decision is morally guided, nor does moral cognition necessarily produce decisions; however, investigations on the computational processes in the brain during decision making are precious for any neurocomputational moral model. Reinforcement learning is the framework of reference, first introduced by Rescorla and Wagner (1972) as a theoretical model of Pavlovian conditioning is a formalization of the problem of how to learn from intermittent positive and negative events in order to improve action selection though time and experience. The agent should act trying to minimize the quantity $\delta(t)$ in the following equation:

$$\delta(t) = r(t+1) + \gamma v_\pi\,(s(t+1)) - v_\pi\,(s(t)) \tag{2}$$

where $t$ is the discrete time sequence of the events, $r$ is the reward, positive or negative, obtained by the environment, which state is an element $s$ of a finite set, $\pi$ is the probabilistic policy of the agent, in choosing an action at the time $t$ in presence of the state of the world $s$, $v$ is the value function, the prediction of rewards, and $\gamma$ is the discount of interest in rewards, as long as they are postponed in time. Solutions to (2) in a model using neuron-like elements were first proposed by Barto and Sutton (1982), Barto et al. (1983). Gradually concepts of reinforcement learning have been fitted into the biology of neuromodulation and forebrain circuits implicated in decision making (Daw et al. 2002; Doya 2002; Dayan 2008; Bullock et al. 2009).

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. Computational reinforcement learning has not yet been studied in order to simulate morality, but a small class of models verge on it, focusing on emotions. As will be explained further bellow, MONE is based on the tight link between morality and emotions. In a broad sense, all reinforcement learning models of basal ganglia and the amygdala deal with the emotional centers of the brain (Levine 2007), but just for a few of them is emotion the matter at hand.

The model GAGE, proposed by Wagar and Thagard (2004), named with reference to the historical case of Phineas Gage, assembles groups of artificial neurons corresponding to the ventromedial prefrontal cortex, the hippocampus, the amygdala, and the nucleus accumbens. It hinges on the somatic-marker idea of Damasio (1994), feelings that have become associated through experience with the predicted long-term outcomes of certain responses to a given situation. GAGE was tested in a simplified version of the Iowa Gambling Task (Bechara et al. 1994), selecting cards from two decks. One can give larger immediate rewards, but a long-term overall loss, the other gives smaller rewards, but a gain in the long term. The model was able to learn to decide for the long-term reward, by virtue of the associations made between the ventromedial prefrontal cortex and the amygdala of the experienced loss.

GAGE implementation of somatic markers was based on Hebbian learning only, without reinforcement learning, which was adopted in ANDREA (Litt et al. 2006, 2008), a model where the orbitofrontal cortex, the dorsolateral prefrontal cortex, and the anterior cingulate cortex interact with basal ganglia and the amygdala. This model was designed to reproduce a well-known phenomenon in economics: the common hypersensitivity to losses over equivalent gains, analyzed in the prospect theory of Kahneman and Tversky (1979). The asymmetric evaluation of gains and losses is simulated in ANDREA at the output of the orbitofrontal cortex, under the effect of the amygdala, conveying emotional arousal. Thagard and Aubie (2008) announced EMOCON, a sophisticated model that incorporates ideas from ANDREA and GAGE, along with simulations of sensorial inputs, that were lacking in both the previous models. One challenging target of this future model is the simulation of emotional consciousness.

The overall architecture of the models of Thagard and his group has several similarities with those of Frank and Claus (2006), Frank et al. (2007), in which the orbitofrontal cortex interacts with the basal ganglia, but more oriented to dichotomic on/off decisions.

## 3 Philsophical foundation of MONE

In general, neurocomputational models target a specific class of known phenomena, trying to reveal new information about the involved mechanism. In modeling morals, one is faced with the problem of establishing a working definition of what morality is. Coming up with a precise definition of morality is exactly one of the main endeavors of moral philosophy. Establishing a clear cut between moral decisions and everyday social problem-solving non-moral decisions, or between

moral norms and social conventions, is not a simple and straightforward task.

Therefore the MONE model, in addition to being based on relevant brain facts, is rooted in a number of theoretical foundations.

### 3.1 Morality is learned emotion

First, at the core of MONE are the emotional brain centers involved in values and decisions, because we embrace the idea that moral cognition is emotional in nature. It is a view within a philosophical tradition that goes back to Hume (1740), among its most authoritative contemporary defenders we have Prinz (2006, 2008a) and Nichols (2004), whose detailed analyses dispense us from exposing the full set of supporting motivations. We limit ourselves to highlighting how the emotional basis of morality has been ascertained in a large number of neuroscientific studies. Moll et al. (2005) reported the remarkable agreement between damage causing deficient emotional engagement, and impairments in moral judgments, as in the case of the ventromedial prefrontal cortex. Dysfunction in these areas is also typical in psychopathy, characterized by poor moral behavior, together with dysfunction in the amygdala (Blair 2007) and in the orbitofrontal cortex (Blair 2010). Moral judgments and emotions seem to coincide in the brain, with structures, in addition to those just mentioned, like the insula, anterior cingulate cortex, the temporal pole, the medial frontal gyrus (Moll et al. 2008). Additional strong evidence comes from studies showing that manipulating emotions can influence moral judgments (Schnall et al. 2008). Cameron et al. (2013) demonstrate that it is even possible, with specific training, to make fine-grained distinctions between emotions that are incidental to the actions being judged versus emotions that are integral to them, discounting inappropriate emotions while making moral judgments.

The second fundamental essence of morality is that it is a learned emotion. Even if emotion is grounded on the same neural equipment evolved in humans for sociality and cognition, evolution falls short of explaining any of our specific moral values. There is nothing like a set of moral rules in our brain (Churchland 2011, pp.163–190), the only biologically based emotional constraints are too abstract and generic to guide specific behavior. On the contrary, all the areas involved in actual morality are highly plastic. The frontal structures belong to the most critical region for learning, storing, and binding social knowledge to contextual elements. In a study with participants aged between 4 and 37 years, Decety et al. (2012) found an age-related increase in activity in the ventromedial prefrontal cortex, as well as increased functional connectivity between this region and the amygdala, in response to dynamic visual stimuli depicting moral transgressions.

The construction of an individual's morality is clearly a product of cultural transmission, reinforced by its emotional appeal, due to the intrinsic content of the belief, the emotional conditioning of the accompanying practices. The experiments executed with MONE in this study belong to this second category. An action, initially pleasant, becomes a moral transgression by the discomfort brought on by adults of the social group. Once the action has been marked with the learned negative emotion, it gets inhibited without the need of external threats. Half of the story is left unanswered in MONE: the reasons why that action is believed to be bad in the social group, thus sanctioned in youngsters. We plan to address this question, at least in part, in future extensions of MONE as multiple agents.

### 3.2 Morality is not a single mechanism

What comes under the label of morality, when examined closely, looks much more like a collection of different patterns of behavior, rather than a monolithic set of beliefs. It comes as a natural consequence of the essence of morals as emotions. There are distinct basic emotions where several classes of moral situations find their place. Stich (2006) has cogently proposed the idea of how dissociated the cognitive pieces that make up morality are, allegorizing morality as a kludge, rather than an elegant machine.

One of the first taxonomies attempted on an empirical basis is the psychological study of Rozin et al. (1999), in which subjects were presented with vignettes that depicted either a clear harm, an instance of disrespect, or a case of something we tend to regard as polluting the body. Subjects were asked to identify the appropriate emotional response. Rozin et.al. proposed a model to explain the results, called CAD for the three emotions contempt, anger, and disgust, but also for the three related classes of moral codes: community, autonomy, and divinity. The CAD model is an important achievement, still very influential in moral cognition, despite the fact that the exact definition of the classes of emotions and the related moral codes have been debated. (Prinz 2008a, pp. 73–75) proposes a first main classification in other- and self-directed moral emotions, with only two basic emotions in the first: anger and disgust. Rozin's contempt is a blending of those two. The reflexive basic moral emotions are guilt and shame. Guilt, which is likely to follow when inflicting physical harm or taking something from a member of the group, is the kind of emotion on which the MONE experiments here described are based.

Recent neurocognitive experiments have confirmed that morality is not a wholly unified faculty, but rather instantiated in partially dissociable neural systems that are engaged differentially depending on the kind of emotion elicited by the moral transgression (Parkinson et al. 2011). Our long-term goal is to gradually extend MONE in order to include

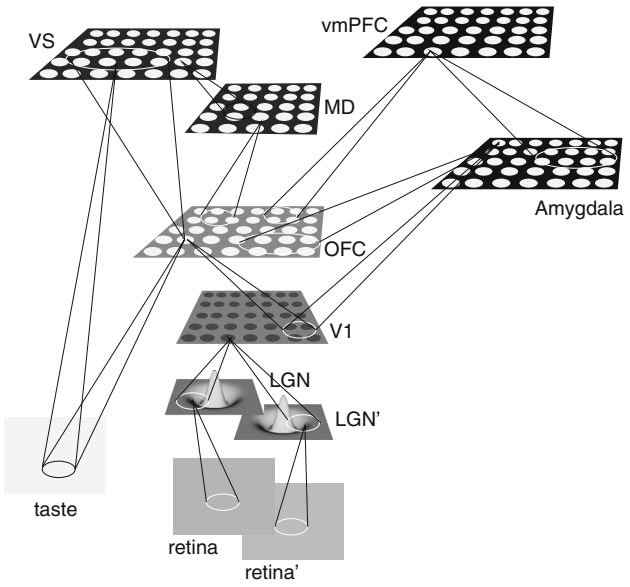


**Fig. 1** Overall scheme of the model composed of LGN (*lateral geniculate nucleus*), V1 (*primary visual area*), OFC (*orbitofrontal cortex*), VS (*ventral striatum*), MD (*medial dorsal nucleus of the thalamus*), Amyg (*amygdala*), vmPFC (*ventromedial prefrontal cortex*)

more classes of moral behaviors, related to other emotions, in addition to those here presented.

## 4 Implementation and neurophysiological justification of MONE

The overall architecture of MONE is shown in Fig. 1. It is composed of a series of sheets with artificial neural units, labeled with the acronym of the brain structure that is supposed to reproduce. All components are described in detail below. Several components derive from previous developments using the same basic artificial neural architecture, in simulating object recognition (Plebe and Domenella 2007), complex pattern responses in visual area V2 (Plebe 2012), language development (Plebe et al. 2010), and a preliminary sketch of decision making in moral context (Plebe 2014). Let us clarify a potential misleading idea in the MONE acronym, by "moral neural engine" we mean nothing like a moral module in the brain. Moral cognition not only is a collection of many different processes, as described in the previous section, even for a single kind of moral situation, it results from activity distributed in most of the brain. As any neurocomputational model of high cognitive functions, MONE is a crude simplification and abstraction of the brain, focusing on just a few areas, that, according to current studies, seem essential to the kind of moral judgment here simulated. In this sense, the neural circuit included in MONE can be dubbed "moral neural engine."

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 and the dopaminergic neurons. 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 model is implemented using the *Topographica* neural simulator (Bednar 2009); most of the components visible in Fig. 1 follow the LISSOM architecture (*Laterally Interconnected Synergetically Self-Organizing Map*) (Sirosh and Miikkulainen 1997), which implements flexible and modifiable lateral connections of both excitatory and inhibitory types.

In a LISSOM sheet of neurons, the activation of each neuron is due to the combination of afferents and excitatory and inhibitory lateral connections. Since the excitatory and inhibitory contributions depend on the activation of neighbor neurons on the same sheet, the computation is recursive in time, in general 10 steps are sufficient for convergence.

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

$$x_i^{(k)} = f\left(\gamma_{\mathrm{A}}\mathbf{a}_i \cdot \mathbf{v}_i \; + \gamma_{\mathrm{E}}\mathbf{e}_i \cdot \mathbf{x}_i^{(k-1)} \; - \gamma_{\mathrm{H}}\mathbf{h}_i \cdot \mathbf{x}_i^{(k-1)}\right) \quad (3)$$

The vector fields $\mathbf{v}_i$, $\mathbf{e}_i$, $\mathbf{x}_i$ are circular areas of radius $r_{\mathrm{A}}$ for afferents, $r_{\mathrm{E}}$ for excitatory connections, $r_{\mathrm{H}}$ for inhibitory connections. The vector $\mathbf{a}_i$ is the receptive field of the unit $i$. Vectors $\mathbf{e}_i$ and $\mathbf{h}_i$ are composed of 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, and $k$ is the time step in the recursive procedure.

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

$$\Delta\mathbf{a}_{r_{\mathrm{A}},i} = \frac{\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}}{\left\|\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}\right\|} - \mathbf{a}_{r_{\mathrm{A}},i}, \quad (4)$$

$$\Delta\mathbf{e}_{r_{\mathrm{E}},i} = \frac{\mathbf{e}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}}{\left\|\mathbf{a}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}\right\|} - \mathbf{e}_{r_{\mathrm{E}},i}, \quad (5)$$

$$\Delta\mathbf{i}_{r_{\mathrm{I}},i} = \frac{\mathbf{i}_{r_{\mathrm{I}},i} + \eta_{\mathrm{I}} x_i \mathbf{x}_{r_{\mathrm{I}},i}}{\left\|\mathbf{i}_{r_{\mathrm{I}},i} + \eta_{\mathrm{I}} x_i \mathbf{x}_{r_{\mathrm{I}},i}\right\|} - \mathbf{i}_{r_{\mathrm{I}},i}, \quad (6)$$

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.

Now all the components will be described, with a justification of their role in morality, from neurophysiological evidence, and their mathematical representation in the model. In the equations, for sake of readability, the following symbols will be used for the subcortical signals:

- ⊙ the output of the LGN at the time when seeing the main scene;
- ◎ the output of the LGN deferred in time, when a possibly angry face will appear;
- ⊡ the taste signals;
- ⊛ the output of the medial dorsal nucleus of the thalamus.

The contents of these signals will be specified in the description of the experiments.

### 4.1 Orbitofrontal cortex

The orbitofrontal cortex is a part of the prefrontal cortex, including Brodmann (1909) areas 12 and 13. It receives a varied assortment of sensorial information from the visual stream, taste, olfactory, auditory, and somatosensory inputs. It is the site of several high level functions, and a summary is given by Rolls (2004).

The visual input is from the ventral stream, concerned with the semantic content of the visual scene (Ungerleider and Mishkin 1982). There are neurons in the orbitofrontal cortex that respond differentially to visual objects depending on their reward association, and one of the primary reinforcements is taste (Rolls et al. 1996). In addition, there is a population of orbitofrontal neurons which respond to faces (Rolls et al. 2006), some of them specifically to facial expressions. The significance of these neurons is likely to be related to the fact that facial expressions convey information that is important in social reinforcement.

The crucial role played by the orbitofrontal cortex in social decision making was discovered through the observation of patients with lesions (Damasio 1994; Bechara et al. 1994). Its specific relevance for morality is controversial. According to Greene and Haidt (2002), this area might perform a general regulative function, in which affective information guides approach and avoidance behavior in both social and non-social contexts. However, the orbitofrontal cortex is almost always involved in moral cognition (Moll et al. 2005). 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 regard to rewards and punishments. Boorman and Noonan (2011) argue that uncertainty about the function of the orbitofrontal cortex in guiding decision making may be due to its internal divisions having distinct functions. Collectively, evidence so far indicates that the orbitofrontal cortex can encode the subjective value of both the expectation and the experience of specific rewards during value-based choices. It is clearly a fundamental function in moral cognition, but in other processes as well.

MONE plays on most of the typical orbitofrontal functions just described: object recognition, their reward association with taste, and the reinforcement of facial expressions. The equation of the activation of a neural unit in the OFC layer is the following:

$$
\begin{aligned}
x^{(\mathrm{OFC})} = f\Big( & \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})} \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})} \cdot \mathbf{v}_{r_{\mathrm{A}}}^{(\mathrm{V1})} \\
& + \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\circledcirc)} \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\circledcirc)} \cdot \mathbf{v}_{r_{\mathrm{A}}}^{(\circledcirc)} \\
& + \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\boxdot)} \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\boxdot)} \cdot \mathbf{v}_{r_{\mathrm{A}}}^{(\boxdot)} \\
& + \gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\circledast)} \mathbf{b}_{r_{\mathrm{B}}}^{(\mathrm{OFC})} \cdot \mathbf{v}_{r_{\mathrm{B}}}^{(\circledast)} \\
& + \gamma_{\mathrm{E}}^{(\mathrm{OFC})} \mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{OFC})} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{OFC})} \\
& - \gamma_{\mathrm{H}}^{(\mathrm{OFC})} \mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{OFC})} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{OFC})} \Big)
\end{aligned}
\tag{7}
$$

which is a specialization of the general Eq. (3). Here, for better readability, the unit index $i$ and time step $k$ have been omitted. There are three sensorial afferents: $\mathbf{v}_{r_{\mathrm{A}}}^{(\mathrm{V1})}$ from the visual cortex V1, $\mathbf{v}_{r_{\mathrm{A}}}^{(\odot)}$ from the lateral geniculate nucleus of the thalamus, and the taste sensorial input $\mathbf{v}_{r_{\mathrm{A}}}^{(\boxdot)}$, each in a sensorial area $r_{\mathrm{A}}$ corresponding to the receptive field of the unit in OFC. A fourth afferent, $\mathbf{v}_{r_{\mathrm{B}}}^{(\circledast)}$, is the diffuse projection from MD, carrying dopamine signaling from the loop that will be described next, and its equation will be given in (12). The visual pathway in the brain travels along many cortical maps, such as V1, V2 and V4, and in MONE it is simplified in a single area, V1, with the following equation:

$$
\begin{aligned}
x^{(\mathrm{V1})} = h\Big( & \gamma_{\mathrm{A}}^{(\mathrm{V1}\leftarrow\odot)} \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{V1}\leftarrow\odot)} \cdot \mathbf{v}_{r_{\mathrm{A}}}^{(\odot)} + \gamma_{\mathrm{E}}^{(\mathrm{V1})} \mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{V1})} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{V1})} \\
& - \gamma_{\mathrm{H}}^{(\mathrm{V1})} \mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{V1})} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{V1})} \Big)
\end{aligned}
\tag{8}
$$

which differs from Eq. (3) in that the nonlinear function $h$ has an adaptive threshold $\theta$, dependent on the average activity of the unit, using:

$$
\theta^{(k)} = \theta^{(k)} + \lambda \left( \bar{x}^{(\mathrm{V1})} - \mu \right) \tag{9}
$$

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:

$$x^{(\odot)} = f\left(\frac{\gamma_O\left(\mathbf{g}_{r_A}^{(\sigma_N)} - \mathbf{g}_{r_A}^{(\sigma_W)}\right)\cdot \mathbf{v}_{r,c}}{\beta + \gamma_S \mathbf{g}_{r_A}^{(\sigma_S)}\cdot \mathbf{x}_S^{(\odot)}}\right) \tag{10}$$

approximating the combined contribution of ganglion cells and LGN with a positive center and negative surround, by differences of two Gaussian $\mathbf{g}^{(\sigma_N)}$ and $\mathbf{g}^{(\sigma_W)}$, 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 $\mathbf{x}_S^{(\odot)}$ are the suppressive connection field of the given unit. It holds $\sigma_N < \sigma_S < \sigma_W$.

The retinal photoreceptors array has size $28 \times 28$, the size of LGN is $24 \times 24$, that of V1 is $22 \times 22$, and that of OFC is $16 \times 16$.

### 4.2 Ventral striatum

The term ventral striatum, VS, was coined by Heimer (1978), to include the nucleus accumbens and the broad continuity in the basal ganglia between the caudate nucleus and putamen. It is at the crossroad of neural networks that treat various aspects of reward processes and motivation. Afferent projections to the VS are derived from three major sources: a massive input from the cerebral cortex, a large input from the thalamus, and a smaller but critical input from the midbrain dopaminergic cells. Cortico-striatal terminals have a topographic organization, with distinct terminal fields from the orbitofrontal cortex, the ventromedial prefrontal cortex, and the anterior cingulate cortex. VS has a direct and reciprocal connection with the dopaminergic neurons located in the substantia nigra pars compacta and the ventral segmental area. This organization is consistent with the findings that diverse striatal areas are activated following reward-related behavioral paradigms (Haber 2011). The dopaminergic neurons are projecting back to the ventral pallidum, and from there to MD, the medial dorsal nucleus of the thalamus, which, in turn, projects to the prefrontal cortex and is the final link in the reward circuit.

This circuit is implemented in MONE by the following two equations:

$$\begin{aligned} x^{(\mathrm{VS})} = f\Big(&\gamma_A^{(\mathrm{VS}\leftarrow\mathrm{OFC})}\mathbf{a}_{r_A}^{(\mathrm{VS}\leftarrow\mathrm{OFC})}\cdot\mathbf{v}_{r_A}^{(\mathrm{OFC})} \\ &+ \gamma_A^{(\mathrm{VS}\leftarrow\boxdot)}\mathbf{a}_{r_A}^{(\mathrm{VS}\leftarrow\boxdot)}\cdot\mathbf{v}_{r_A}^{(\boxdot)} \\ &+ \gamma_E^{(\mathrm{VS})}\mathbf{e}_{r_E}^{(\mathrm{VS})}\cdot\mathbf{x}_{r_E}^{(\mathrm{VS})} - \gamma_H^{(\mathrm{VS})}\mathbf{h}_{r_H}^{(\mathrm{VS})}\cdot\mathbf{x}_{r_H}^{(\mathrm{VS})}\Big) \end{aligned} \tag{11}$$

$$x^{(\circledast)} = f\left(\gamma_A^{(\circledast\leftarrow\mathrm{VS})}\mathbf{a}_{r_A}^{(\circledast\leftarrow\mathrm{VS})}\cdot\mathbf{v}_{r_A}^{(\mathrm{VS})}\right) \tag{12}$$

The afferent signals $\mathbf{v}^{(\mathrm{OFC})}$ come from Eq. (7), and $\mathbf{v}^{(\boxdot)}$ is the taste signal. The output $x^{(\circledast)}$ computed in (12) will close the loop into the prefrontal cortex with Eq. (7). In this equation, there is a parameter, $\gamma_B^{(\mathrm{OFC}\leftarrow\circledast)}$, which will be used in a special way during the experiments with MONE. It is a global modulatory factor of the amount of dopamine signaling for gustatory reward, and therefore it is the most suitable parameter for simulating hunger states. In fact, despite decades of investigation, considerable differences of opinion still exist about how food intake is controlled by the brain (Woods and Stricker 1999), and it is out of the scope of this work to model this control in detail. The purpose here is to modulate in a plausible way the internal drive to eat, which will be taken into account among other factors in the decision process of the model.

The VS map has size $8 \times 8$, and MD has size $4 \times 4$.

### 4.3 Ventromedial prefrontal cortex

Since the early studies on patients with lesions in the ventromedial prefrontal cortex, vmPFC, this region has been shown to play a crucial role in emotion regulation and social decision making (Bechara et al. 1994; Damasio 1994). Damage to vmPFC can produce devastating impairments in higher-order behavioral guidance, social conduct, and easily induce a subject to choosing highly aversive personal actions. Recently, researchers have begun to investigate the specificity of vmPFC with respect to other prefrontal components, like OFC, of the dorsolateral frontal cortex, which are also involved in emotion-based decision making, as discussed above. According to Hernandez et al. (2009), vmPFC stands out as the heart of neural machinery involved in emotional intelligence, the ability to reliably regulate and utilize emotional information in evaluating choices. 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). Boorman and Noonan (2011) found that vmPFC encoded the relative chosen value: the chosen expected value relative to the unchosen expected value, rather than the absolute value of a choice. In this way, vmPFC may actively compare the available options competing for choice.

According to Greene and Haidt (2002), this area is involved in guiding approach and avoidance behavior in many contexts and is not specific for moral judgment. On the other hand, Moll et al. (2005) argue that the vmPFC is implicated in representing social and emotional structured event complexes (SEC), which are situational knowledge abstracted across events, and the temporal organization of events. Although this framework is relevant for social behavior in general, it has clear implications for moral cognition. Recently Decety et al. (2012), in a neurodevelopmental study, found that seeing intentional harm to people results in stronger activation in vmPFC in adults than in younger children. In addition, an age-related change has been found in the

functional integration between vmPFC and amygdala, such that the older participants showed significant coactivation in these regions when attending to scenarios with intentional harm.

vmPFC is implemented in MONE using the standard Eq. (3), as follows:

$$x^{(\mathrm{vFC})} = f\Big(\gamma_{\mathrm{A}}^{(\mathrm{vFC}\leftarrow\mathrm{OFC})}\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{vFC}\leftarrow\mathrm{OFC})}\cdot\mathbf{v}_{r_{\mathrm{A}}}^{(\mathrm{OFC})} + \gamma_{\mathrm{A}}^{(\mathrm{vFC}\leftarrow\mathrm{Amy})}\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{vFC}\leftarrow\mathrm{Amy})}\cdot\mathbf{v}_{r_{\mathrm{A}}}^{(\mathrm{Amy})} + \gamma_{\mathrm{E}}^{(\mathrm{vFC})}\mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{vFC})}\cdot\mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{vFC})} - \gamma_{\mathrm{H}}^{(\mathrm{vFC})}\mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{vFC})}\cdot\mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{vFC})}\Big) \tag{13}$$

The afferent signals $\mathbf{v}^{(\mathrm{OFC})}$ come from Eq. (7), while $\mathbf{v}^{(\mathrm{Amy})}$, the connection from Amygdala, is given from Eq. (14). The size of the vmPFC map in the model is $12 \times 12$.

### 4.4 Amygdala

The crucial connection between vmPFC and the amygdala in moral cognition has been emphasized by other researchers too. Blair (2007) alleges that learning the basics of care-based morality relies on the crucial role of the amygdala in stimulus-reinforcement learning, and in turn this learning enables representations of conditioned stimuli within vmPFC to be linked to emotional responses.

The amygdala, located deep and medially within the temporal lobes, was first recognized as primary mediator of negative emotions, especially fear, and responsible for learning associations that signal a situation as fearful LeDoux (2000). Further studies have shown that emotional responses mediated by the amygdala cover a wide range, such as anger and sadness, and not limited to negative emotions only. In short, the amygdala enables the individual to learn the goodness and badness of objects and actions (Blair 2007). Learning in the amygdala involves the intersection of pathways transmitting sensorial information, from both cortical sensory areas and direct thalamic connections.

The involvement of the amygdala in the recognition of facial expressions is well documented, with different kinds of expressions clustered in different subregions, and with the strongest activation in response to direct-gazing angry faces (Boll et al. 2011). Pegna et al. (2004) even reported on a subject who, after bilateral destruction of his visual cortices and ensuing cortical blindness, could nevertheless correctly guess the type of emotional facial expression being displayed, by activation of his right amygdala.

In MONE this specialized recognition is exploited, by using faces with angry expressions in the experiments. The activation of units in the artificial amygdala component is given by:

$$x^{(\mathrm{Amy})} = f\Big(\gamma_{\mathrm{A}}^{(\mathrm{Amy}\leftarrow\mathrm{OFC})}\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{Amy}\leftarrow\mathrm{OFC})}\cdot\mathbf{v}_{r_{\mathrm{A}}}^{(\mathrm{OFC})} + \gamma_{\mathrm{A}}^{(\mathrm{Amy}\leftarrow\circledcirc)}\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{Amy}\leftarrow\circledcirc)}\cdot\mathbf{v}_{r_{\mathrm{A}}}^{(\circledcirc)} + \gamma_{\mathrm{E}}^{(\mathrm{Amy})}\mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{Amy})}\cdot\mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{Amy})} - \gamma_{\mathrm{H}}^{(\mathrm{Amy})}\mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{Amy})}\cdot\mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{Amy})}\Big) \tag{14}$$

The afferent signals $\mathbf{v}^{(\mathrm{OFC})}$ come from Eq. (7), while $\mathbf{v}^{(\circledcirc)}$ 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 Eq. (14) will loop inside the vmPFC by Eq. (13). The size of the amygdala map in the model is $8 \times 8$.

### 4.5 Comparison with other models

A true comparison with alternatives is not possible, because MONE is the first neurocomputational model that simulates moral behavior; however, as mentioned in Sect. 2.3, there are existing models that simulate, in non-moral experiments, mechanisms shared by MONE. The model GAGE (Wagar and Thagard 2004) includes the area vmPFC, the ventral striatum (called nucleus accumbens) like MONE, with the addition of the hippocampus. The successor model ANDREA (Litt et al. 2006, 2008) lacks the hippocampus, but the frontal areas are simulated in more detail, with separate orbitofrontal, dorsolateral, and anterior cingulate cortices. The models of Frank and Claus (2006), Frank et al. (2007), on the other hand, have less details in the frontal component, with just a simulated orbitofrontal area, but more detail in the basal ganglia, with separated roles for the globus pallidus, the substantia nigra, and two subpopulations of cells in the striatum. The main difference in all these models is in the dimension, and related complexity, of the signals processed by the components. In GAGE and ANDREA, the only input to the system is a unidimensional signal that goes positive in the case of rewards, and negative for potential losses. Also in the models of Frank and Claus (2006), Frank et al. (2007), the input is highly abstract, coding for two possible alternate inputs only. The lack of a rich sensorial interface makes all those models unsuited to be expanded for simulating morality, because even the simplest form of moral contingency requires the recognition of objects and events in the microworld, where the model is embedded, and where it should learn to behave morally. In MONE this is made possible by including visual and taste senses, consequently designing the size of all brain components large enough to process the detailed representations of the external world. Even if the moral contingency is essential, it has been designed in a way respecting biological plausibility as much as possible; therefore, the taste reward has been chosen because it is a direct afferent to the orbitofrontal cortex, and

**Fig. 2** The visual inputs of the model. From the *left to the right*: an edible object, possibly an apple, a +-shaped and a *X*-shaped neutral object, the edible object in the forbidden area (the *bottom right* quadrant of the scene), synthetic elongated blobs used in the early visual development, the sad and angry schematic face

the negative reaction through the angry face has been chosen to exploit the feature of the amygdala to recognize face emotions.

On the other hand, in the other model there are more details on the decision-making components, especially in the simulation of the basal ganglia, these details are not the focus of the analysis in MONE, which uses a simplified and more compact simulation of those areas.

## 5 Experiments and results

The artificial moral brain architecture just described is exposed to a series of situations that simulate highly simplified contexts, and MONE can choose between different actions. 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."

The main input to the model is a visual scene, where three types of objects can appear, at random positions. There is a kind of object with an *X* shape and another with a + shape, which are neutral for the individual fitted with the MONE mini-brain. Only one object is edible, the one with a spherical shape, and it might be an apple, an example of a renowned sacred prohibition. Examples are shown in the first and fourth positions from the left in Fig. 2.

Our artificial subject is unfamiliar with the objects, and she can realize how pleasant fruits are to eat, thanks to her taste perception. This sensorial input is simply a matrix $2 \times 2$, in which the ratio of the upper row to the lower row signal represents how pleasant the taste is. To follow the biblical resemblance, there is an Eden, in this case, simply one quadrant of the overall scene. It is forbidden to eat apples in the bottom right quadrant. In a more profane context, fruits in this quadrant may belong to a member of the social group, and to collect these fruits would be a violation of her/his property, which 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

The model can choose between two possible behaviors: collect and eat an object, or refrain from doing it. Of course it is unnecessary to simulate a motor action, just its selection suffices, which is coded in the vmPFC component. In order to characterize the ensemble activation pattern of the vmPFC neurons, and decode the chosen action, a population code method is applied. The overall population is clustered according to those neurons that were active in response to different classes of objects compared to those that were not responsive, and mathematical details are in (Plebe et al. 2011; Plebe 2014). The decoded representations in vmPFC were used to determine which choice MONE had opted for in a given experiment.

### 5.1 Experiment 1

The MONE-equipped subject passes through sequential phases of development, in this first experiment the following two:

1. an early stage of formation of the visual system, with the elongated patterns, shown in Fig. 2 (second from the right), as the only type of stimuli;
2. the good food recognition stage, in which the stimuli are the three types of objects, in all possible positions, and their taste;

Learning is always ruled by Eqs. (4), (5), and (6) applied to the relevant connections, and in the first stage are those of Eq. (8) only. The development of V1 set up the main systems of organization in the primary visual cortex, with arrangement of orientation tuned neurons, similar to that described in (XX,YY2).

In the second stage, the OFC, VS, MD, and vmPFC areas of the model are plastic and learn their connections of Eqs. (7), (11), (12), and (13). 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 results in the OFC map are analyzed using the population code procedure previously described, and the populations of neurons coding for the three kind of objects are shown in Fig. 3. Coding for the edible objects is more

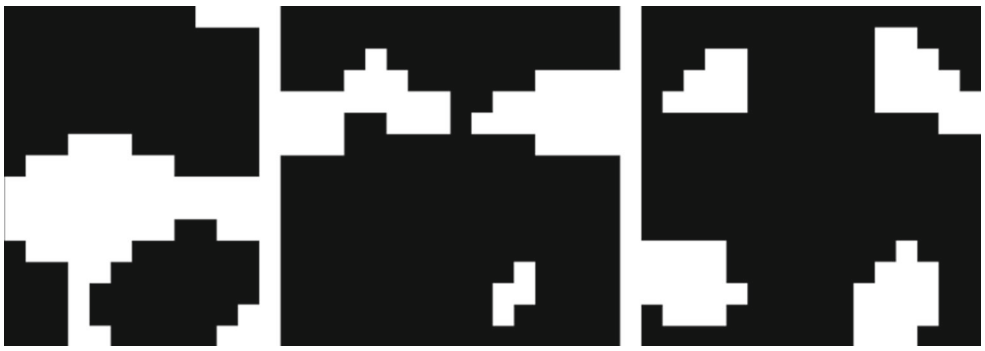

**Fig. 3** Representation of the three possible objects of the small simulated world, in the OFC model map, by population coding. From *left to the right*, the neurons coding for the edible object, the $X$-shaped objects, and the $+$-shaped objects

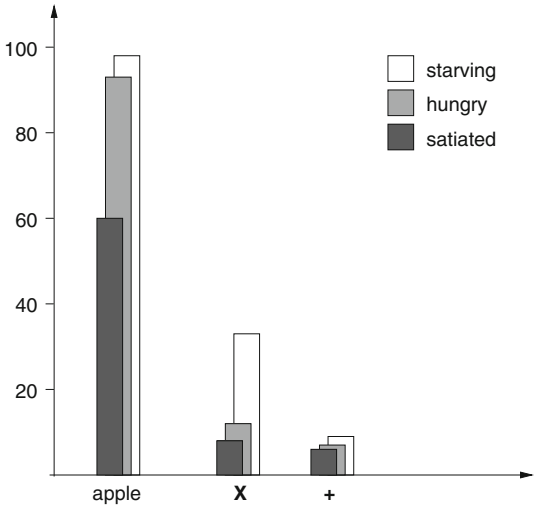


**Fig. 4** Fraction of grasping actions selected by the vmPFC model map, depending on the object seen. The model is tested with three different conditions of simulated hunger

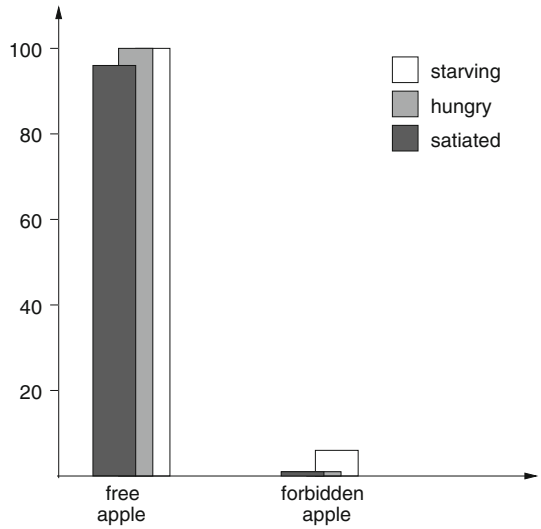


**Fig. 5** Fraction of grasping actions selected by the vmPFC model map, depending on the position of the edible object: free or in the forbidden area. The model is tested with three different conditions of simulated hunger

compact and exploits a larger part of the neural map. The population coding for the $X$-shaped and $+$-shaped objects are more scattered and partially overlapped.

The coding in vmPFC model map is the decision made to grasp or not to grasp the object, and the percentage of decision to grasp, for each type of object, is shown in Fig. 4. The model is tested using three different levels of the parameter $\gamma_B^{(OFC \leftarrow \circledast)}$ of Eq. (7) that simulate the internal hunger drive. When the object is an apple, grasping is always the prevailing choice that drops to 60 % only in satiated condition. Occurrence of grasping is instead meaningless for the other objects, except while starving. In this situation, the model decides to grasp $X$-shaped objects about 30 % times and the $+$-shaped objects about 10 % times, even if these objects do not reward.

### 5.2 Experiment 2

This second experiment starts with the model at the final level of training, reached at the end of experiment 1. Now it experiences a third stage, 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 Eq. (14), 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. 5 there are the percentages of decisions to grasp an object, 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 strong the inhibition to grasp the edible objects

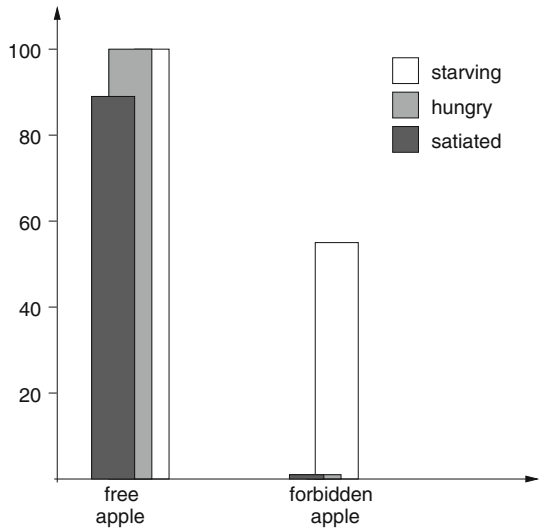


**Fig. 6** Fraction of grasping actions selected by the vmPFC model map, as in Fig. 5, when the model amygdala is lesioned

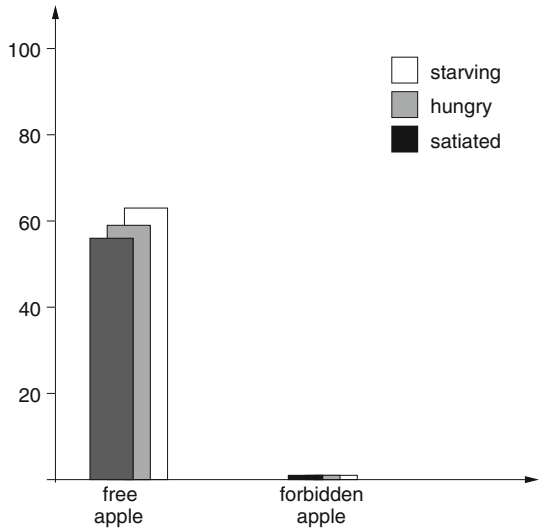


**Fig. 7** Fraction of grasping actions selected by the vmPFC model map, as in Fig. 5, when the model vmPFC is lesioned

is when placed in the forbidden sector. In both the conditions of normal hunger or satiation, no one grasping decision is issued for forbidden apples, while the same fruit in the free territory is grasped 96,% of the times when satiated, and 100 % when hungry. Only under the extreme starving condition are there limited cases of transgression, 0.6 % of the times. It can be claimed that the model has learned a moral rule, as an imperative inhibition to perform certain actions.

### 5.3 Experiment 3

The third experiment is a follow-up of experiment 2, in which the model that has learned that it is bad to steal apples becomes lesioned.

Figure 6 shows the results of the model, when, after the same training of experiment 2, in the amygdala map 25 % of the units are switched off. There are no significant differences in the condition of normal hunger or satiation, just a decreased tendency to grasp free apples when satiated, 89 % against 96 % of the healthy model. The main difference is in the starving condition. Now the model decides to grasp forbidden apples in 55 % of the cases. This behavior closely resembles that of individuals with psychopathy, who show significantly less of a moral-conventional distinction than do healthy individuals. The moral norms are not neglected, but transgressions are permissible in contingent situations where their reward is especially high. The imperative inhibition of the moral norm is loosened.

Figure 7 shows the results of the model when, again after the same training of experiment 2, the vmPFC is lesioned, by switching off 25 % of the units in the map. In this case, the lesion has no effect on the moral norm; on the contrary, it is even more effective, inhibiting the decision to grasp forbidden apples even when starving. The main consequence is in an overall reduction in deciding to grasp at any time. In front of free apples, the percentage of grasping decisions is around 60 % for every hunger condition, with small modulatory effect. The vmPFC lesion with the intact amygdala is likely to induce a more apathetic behavior in the model. In the MONE model, intact morality is clearly heavily dependent on an intact amygdala, but the correct functioning of the decision process, taking into account both internal drives and moral rules, depends on the vmPFC as well.

These results match well with neurocognitive human data. Overall evidence that dysfunctions to the amygdala and ventromedial prefrontal cortex are associated with psychopathy, and in particular can compromise moral reasoning, is reviewed in Blair (2007). In a specific study on fear conditioning, supported by the amygdala circuit as in MONE, psychopaths displayed no significant activity, and this dissociation of emotional processing is supposed to be the neural basis of the lack of anticipation of aversive events in criminal psychopaths (Birbaumer et al. 2005). Even closer to the simulated settings of MONE is the study by Marsh et al. (2008) that demonstrated a deficit in amygdala responses to facial fear expressions, in youths with callous-unemotional traits and disruptive behavior disorders.

Due to its crucial role in decision making in general, vmPFC lesions have of course a wide range of negative behavior effects (Fuster 2008), most of which are out of the scope of the simulations with MONE. There are indeed specific studies with data that are consistent with the model results. In a task where subjects are requested to indicate their preference for stimuli in categories including fruits, vegetables, colors, landscapes, and puppies (Henri-Bhargavas et al. 2012), subjects with damage in the ventromedial frontal

lobe were significantly more erratic than control subjects in perceptual judgments. Evidences for apathetic behavior following vmPFC lesions have also been reported (Barrash et al. 2000; Tekin and Cummings 2002; Fellows and Farah 2005), and explained as a deficit in the future time perspective, typical of vmPFC damage.

### 5.4 Psychological interpretation of the experiments

The outcomes of the experiments here described are prone to being interpreted in terms of human moral behavior. The first remark is that no predefined moral mechanism is necessary for an agent to behave correctly, at least in the case of stealing. The moral norm against theft is learned, by means of general-domain brain components: selectivity to face expressions, association of negative emotions. This result is in support of moral theories that do not require an innate endowment of moral values and rules (Prinz 2008b, 2009).

In addition, the model suggests how mechanisms which are not moral specific can be combined in a way that qualifies a decision as plainly moral. The decision of the model to collect apples and avoid other objects in experiment 1 is the result of a purely utilitarian computation, taking into account the reward of the chosen actions, without moral engagement. In experiment 2, in deciding whether to steal apples or not, in addition to the utilitarian computation, there is the emotional component at play, aroused by the negative feeling coded by the amygdala. This result diverges from theories that ground moral decisions on cognitive computations discarding the role of emotion (Mikhail 2009), while supporting emotion-based theories (Nichols 2004; Prinz 2006). It might be argued that a similar mechanism by which the moral norm is acquired in the model may well work for non-moral norm acquisition. Caregivers may display an angry face when a child violates etiquette or good manner norms as well. In fact, the model would not be able to make a difference, if the route to learning a norm is the emotional reaction to an angry face, but this is exactly the case when it is argued that the conventional norm does have a moral feature, due exactly to the way it has been taught (Kelly et al. 2007).

As with any computational approach, the above speculations can be challenged for not being derived directly from humans, which is the only reliable way to gain empirical knowledge. As a matter of fact, at a closer look, experimental moral psychology with humans is exposed to a number of inconveniences which computational models are immune to. The most serious weakness is in the setting of realistic scenarios, in which the judgments of the subjects derive from genuine brain computations normally involved in moral behavior. In the typical experimental setting in moral psychology, subjects are presented with a narration of a scenario, in which they should imagine to engage. This engagement suffers from drawbacks common to thought experiments, like those identified by (Gendler 2007, p.69): "by presenting content in a suitably concrete or abstract way, thought experiments recruit representational schemas that were otherwise inactive, thereby evoking responses that may run counter to those evoked by alternative presentations of relevantly similar content"; and "exactly because they recruit heretofore uninvolved processing mechanisms, thought experiments can be expected to produce responses to the target material that remain in disequilibrium with responses to the same material under alternative presentations".

The situation is even worse in the case of moral scenarios, where evoking representational schemata to imagine one's own engagement may interfere dramatically with direct moral computation. Experiments like the classical trolley problem suffer the lack of *ecological validity* (Casebeer and Churchland 2003), loosing, among others, the "hot" feature of moral behavior: imagining becoming *angry* under certain circumstances is much different from really being angry. There is large evidence on how responses to moral problem tasks can change drastically depending on how the task is proposed. Gold et al. (2014) found differences in moral judgments between outlandish and realistic hypothetical scenarios, and between hypothetical scenarios and the same scenarios operationalized in real-life. Even the level of details given in a moral judgment task is relevant, in an experiment Nichols and Knobe (2007) found that only 34 % of subjects believed an agent was fully morally responsible of a shameful act, while 72 % responded that he was, when the same scenario was described in more detail. This limitation does not exist within a computational model like MONE, where the responses of the artificial agent directly derive from its moral machine, under the given experimental circumstances, and the model need not reconstruct by imagination its own involvement.

## 6 Conclusions

We have described MOral Neural Engine (MONE), a first attempt to simulate moral cognition in a neurocomputational model. Despite its ambitious name, the model in the implementation here presented 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, we believe, is a collection of several, partially dissociated mechanisms, a model must necessarily, at least in its first implementation, choose a specific one to target. Still, it could be argued that stealing is not the best prototype, for example physical harm is often held to be the most compelling moral violation. Second, even in the 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 a 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 MONE, like the orbitofrontal cortex and the amygdala.

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 made.

We think this result picks up on one core aspect of morality: the emergence of a norm, the answer to the famous dilemma posed by Socrates to Euthyphro: whether the actions are pious because the gods approve of them or whether the gods approve of certain actions because they are pious. As well pointed out by (Prinz 2008a, pp.118–119), the intuitive appeal of the latter option can be well accommodated inside a sensibility theory, by a feedback process, in which moral emotions induce moral norms, which simultaneously reinforce moral emotions. This is exactly what happens in the model.

The model predicts how obeying this norm is an imperative that supersedes other internal drives, like hunger, up to a certain extent. This threshold is dramatically lowered when simulating lesions in the ventromedial cortex and in the amygdala.

We believe that the neurocomputational approach is an additional important path in pursuing a better understanding of morals, and MONE, despite the limitations here discussed, is a valid starting point.

## References


Barrash J, Tranel D, Anderson SW (2000) Acquired personality disturbances associated with bilateral damage to the ventromedial prefrontal region. Dev Neuropsychol 18:355–381

Barto A, Sutton R (1982) Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. Behav Brain Sci 4:221–234

Barto A, Sutton R, Anderson C (1983) Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybern 13:834–846

Bechara A, Damasio AR, Damasio HR, Anderson SW (1994) Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 50:7–15

Bednar JA (2009) Topographica: building and analyzing map-level simulations from python, C/C++, MATLAB, NEST, or NEURON components. Front Neuroinformatics 3:8

Bekoff M (2001) Social play behavior. cooperation, fairness, trust, and the evolution of morality. J Conscious Stud 9:81–90

Bekoff M, Pierce J (2009) Wild justice: the moral lives of animals. University of Chicago Press, Chicago

Birbaumer N, Veit R, Lotze M, Erb M, Hermann C, Grodd W, Flor H (2005) Deficient fear conditioning in psychopathy: a functional magnetic resonance imaging study. Arch Gen Psychiatry 62:799–805

Blair J (2007) The amygdala and ventromedial prefrontal cortex in morality and psychopathy. Trends Cogn Sci 11:387–392

Blair J (2010) Neuroimaging of psychopathy and antisocial behavior: a targeted review. Curr Psychiatry Rep 12:76–82

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

Boorman ED, Noonan MP (2011) Contributions of ventromedial prefrontal and frontal polar cortex to reinforcement learning and value-based choice. In: Mars RB, Sallet J, Rushworth MFS, Yeung N (eds) Neural basis of motivational and cognitive control. MIT Press, Cambridge, pp 55–74

Brodmann K (1909) Vergleichende Lokalisationslehre der Grosshirmrinde. Barth, Leipzig

Bullock D, Tan CO, John YJ (2009) Computational perspectives on forebrain microcircuits implicated in reinforcement learning, action selection, and cognitive control. Neural Netw 22:757–765

Cameron D, Payne K, Doris JM (2013) Morality in high definition: emotion differentiation calibrates the influence of incidental disgust on moral judgments. J Exp Soc Psychol 49:719–725

Casebeer WD, Churchland PS (2003) The neural mechanisms of moral cognition: a multiple-aspect approach to moral judgment and decision-making. Biol Philos 18:169–194

Chagnon NA (1974) Studying the Yanomamö. Holt, Rinehart and Winston, New York

Churchland PS (2011) Braintrust: what neuroscience tells us about morality. Princeton University Press, Princeton

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

Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. Neural Netw 15:603–616

Dayan P (2008) Connections between computational and neurobiological perspectives on decision making. Cogn Affect Behav Neurosci 8:429–453

Decety J, Michalska KJ, Kinzler KD (2012) The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. Cereb Cortex 22:209–220

Doris JM, Stich SP (2005) As a matter of fact: empirical perspectives on ethics. In: Smith M,Jackson F (ed) The oxford handbook of contemporary philosophy. Oxford University Press, Oxford

Douglas-Hamilton I, Bhalla S, Wittemyer G, Vollrath F (2006) Behavioural reactions of elephants towards a dying and deceased matriarch. Appl Anim Behav Sci 100:87–102

Doya K (2002) Metalearning and neuromodulation. Neural Netw 15:495–506

Fellows LK, Farah MJ (2005) Dissociable elements of human foresight: a role for the ventromedial frontal lobes in framing the future, but not in discounting future rewards. Neuropsychologia 43:1214–1221

Frank MJ, Claus ED (2006) Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol Rev 113:300–326

Frank MJ, Scheres A, Sherman SJ (2007) Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. Philos Trans R Soc B 362:1641–1654

Frankena W (1939) The naturalistic fallacy. Mind 48:464–477

Fuster JM (2008) The prefrontal cortex, 4th edn. Academic Press, New York

Gendler TS (2007) Philosophical thought experiments, intuitions, and cognitive equilibrium. Midwest Stud Philos 31:68–89

Gläscher J, Hampton AN, O'Doherty JP (2009) Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cereb Cortex 19:483–495
Gold N, Pulford BD, Colman AM (2014) The outlandish, the realistic, and the real: contextual manipulation and agent role effects in trolley problems. Front Psychol 5:35
Greene JD, Haidt J (2002) How (and where) does moral judgment work? Trends Cogn Sci 6:517–523
Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) fMRI investigation of emotional engagement in moral judgment. Science 293:2105–2108
Gunkel D (2012) The machine question: critical perspectives on AI, robots, and ethics. MIT Press, Cambridge
Haber SN (2011) Neural circuits of reward and decision making: integrative networks across corticobasal ganglia loops. In: Mars RB, Sallet J, Rushworth MFS, Yeung N (eds) Neural basis of motivational and cognitive control. MIT Press, Cambridge, pp 22–35
Hamilton WD (1963) The evolution of altruistic behaviour. Am Nat 97:354–356
Hamilton WD (1964) The genetical evolution of social behaviour. J Theor Biol 7:1–52
Heimer L (1978) The olfactory cortex and the ventral striatum. In: Livingston K, Hornykiewicz O (eds) Limbic mechanisms. The continuing evolution of the limbic system concept. Plenum Press, New York
Henri-Bhargavas A, Simioni A, Fellows LK (2012) Ventromedial frontal lobe damage disrupts the accuracy, but not the speed, of value-based preference judgments. Neuropsychologia 50:1536–1542
Hernandez M, Denburg NL, Tranel D (2009) A neuropsychological perspective on the role of the prefrontal cortex in reward processing and decision-making. In: Dreher JC, Tremblay L (eds) Handbook of reward and decision making. Academic Press, New York, pp 291–306
Hume D (1740) A treatise of human nature, vol 3. Thomas Longman, London
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decisions under risk. Econometrica 47:313–327
Kelly D, Stich S, Haley KJ, Eng SJ, Fessler DMT (2007) Harm, affect, and the moral/conventional distinction. Minds Lang 22:117–131
LeDoux JE (2000) Emotion circuits in the brain. Annu Rev Neurosci 23:155–184
Lehmann L, Keller L (2006) The evolution of cooperation and altruism: a general framework and a classification of models. J Evol Biol 19:1426–1436
Levine D (2007) Neural network modeling of emotion. Phys Life Rev 4:37–63
Litt A, Eliasmith C, Thagard P (2006) Why losses loom larger than gains: Modeling neural mechanisms of cognitive-affective interaction. In: Sun R, Miyake N (eds) Proceedings of the 28th annual meeting of the cognitive science society, Lawrence Erlbaum Associates, Mahwah, pp 495–500
Litt A, Eliasmith C, Thagard P (2008) Neural affective decision theory: choices, brains, and emotions. Cogn Syst Res 9:252–273
Marsh AA, Finger EC, Mitchell DG, Reid ME, Sims C, Kosson DS, Towbin KE, Leibenluft E, Pine DS, Blair R (2008) Reduced amygdala response to fearful expressions in children and adolescents with callous-unemotional traits and disruptive behavior disorders. Am J Psychiatry 165:712–720
Mikhail J (2009) Moral grammar and intuitive jurisprudence: a formal model of unconscious moral and legal knowledge. In: Bartels D, Bauman C, Skitka L, Medin D (eds) Moral judgment and decision making. Academic Press, New York
Moll J, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J (2005) The neural basis of human moral cognition. Nat Rev Neurosci 6:799–809
Moll J, de Oliveira-Souza R, Zahn R, Grafman J (2008) The cognitive neuroscience of moral emotions. In: Sinnott-Armstrong W (ed) Moral psychology, vol 3: the neuroscience of morality: emotion, brain disorders, and development. MIT Press, Cambridge, pp 1–18
Moore GE (1903) Principia ethica. Cambridge University Press, Cambridge
Nichols S (2004) Sentimental rules: on the natural foundations of moral judgment. Oxford University Press, Oxford
Nichols S, Knobe J (2007) Moral responsibility and determinism: the cognitive science of folk intuitions. Nous 41:663–685
Parkinson C, Sinnott-Armstrong W, Koralus PE, Mendelovici A, McGeer V, Wheatley T (2011) Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. J Cogn Neurosci 23:3162–3180
Pegna A, Khateb A, Lazeyras F, Seghier M (2004) Discriminating emotional faces without primary visual cortices involves the right amygdala. Nat Neurosci 8:24–25
Plebe A (2012) A model of the response of visual area V2 to combinations of orientations. Netw Comput Neural Syst 23:105–122
Plebe A (2014) A neural model of moral decisions. In: Madani K, Filipe J (eds) Proceedings of NCTA 2014—international conference on neural computation theory and applications, Scitepress, pp 111–118
Plebe A, Domenella RG (2007) Object recognition by artificial cortical maps. Neural Netw 20:763–780
Plebe A, Mazzone M, De La Cruz VM (2010) First words learning: a cortical model. Cogn Comput 2:217–229
Plebe A, Mazzone M, De La Cruz VM (2011) A biologically inspired neural model of vision-language integration. Neural Netw World 21:227–249
Prehn K, Heekeren HR (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 J, Schrijver JD, Vanneste S, Braeckman J (eds) The moral brain essays on the evolutionary and neuroscientific aspects of morality. Springer, Berlin
Prinz J (2006) The emotional basis of moral judgments. Philos Explor 9:29–43
Prinz J (2008a) The emotional construction of morals. Oxford University Press, Oxford
Prinz J (2009) Against moral nativism. In: Murphy D, Bishop M (eds) Stephen stich and his critics. Basil Blackwell, Oxford, pp 167–189
Prinz JJ (2008b) Is morality innate? In: Sinnott-Armstrong W (ed) Moral psychology, vol 2, the cognitive science morality: intuition and diversity. MIT Press, Cambridge, pp 367–406
Rescorla RA, Wagner AR (1972) A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF (eds) Classical conditioning II: current theory and research. Appleton Century Crofts, New York, pp 64–99
Rolls E (2004) The functions of the orbitofrontal cortex. Biol Cybern 55:11–29
Rolls E, Critchley H, Mason R, Wakeman EA (1996) Orbitofrontal cortex neurons: role in olfactory and visual association learning. J Neurophysiol 75:1970–1981
Rolls E, Critchley H, Browning AS, Inoue K (2006) Face-selective and auditory neurons in the primate orbitofrontal cortex. Exp Brain Res 170:74–87
Rozin P, Lowery L, Haidt J, Imada S (1999) The cad triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). J Pers Soc Psychol 76:574–586

Schnall S, Haidt J, Clore GL, Jordan AH (2008) Disgust as embodied moral judgment. Pers Soc Psychol Bull 34:1096–1109

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

Stevens JLR, Law JS, Antolik J, Bednar JA (2013) Mechanisms for stable, robust, and adaptive development of orientation maps in the primary visual cortex. JNS 33:15747–15766

Stich SP (2006) Is morality an elegant machine or a kludge? J Cogn Cult 6:181–189

Tekin S, Cummings JL (2002) Frontal-subcortical neuronal circuits and clinical neuropsychiatry. J Psychosom Res 53:647–654

Thagard P, Aubie B (2008) Emotional consciousness: a neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. Conscious Cogn 17:811–834

Turnbull CM (1973) The mountain people. Simon and Schuster, New York

Ungerleider L, Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA, Mansfield RJW (eds) Analysis of visual behavior. MIT Press, Cambridge, pp 549–586

Wagar BM, Thagard P (2004) Spiking phineas gage: a neurocomputational theory of cognitive-affective integration in decision making. Psychol Rev 111:67–79

Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford

Woods S, Stricker E (1999) Food intake and metabolism. In: Zigmond M, Bloom F, Landis S, Roberts J, Squire L (eds) Fundamental neuroscience. Academic Press, New York