Biologically Plausible Modelling of Morality

Alessio Plebe

Department of Cognitive Science – University of Messina
v. Concezione 8, Messina, Italy
aplebe@unime.it

Abstract. Neural computation has an extraordinarily influential role in the study of several human capacities and behavior. It has been the dominant approach in the vision science of the last half century, and it is currently one of the fundamental methods of investigation for several higher cognitive functions. Yet, no neurocomputational models have been proposed for morality. Computational modeling in general has been scarcely pursued in morality, and existent non-neural attempts have failed to account for the mental processes involved during moral judgments. In this paper we argue that in the past decade the situation has evolved in a way that subverted the insufficient knowledge on the basic organization of moral cognition in brain circuits, making the project of modeling morality in neurocomputational terms feasible. We will sketch an original architecture that combines reinforcement learning and Hebbian learning, aimed at simulating forms of moral behavior in a simple artificial context.

Keywords: moral cognition; orbitofrontal cortex; amygdala

1 Introduction

Neural computation has an extraordinarily influential role in the study of several human capacities and behavior, however no neurocomputational models have been proposed yet for morality, a failure clearly due to the lack of empirical brain information.

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

In this paper we argue that in the past decade the situation has evolved in a way that makes the project of modeling morality in neurocomputational terms feasible. Even if there are no moral models yet, existing developments in simulating emotional responses and decision making are already offering important frameworks that we think can support the project of modeling morality. The existing models deemed closer to what pertains to morality will be shortly reviewed. We will also sketch an original architecture that combines reinforcement learning and Hebbian learning, aimed at simulating forms of moral behavior in a simple artificial context, and show its few preliminary results.

2 Other approaches to moral computing

Two computational accounts of morality, different from neurocomputation, will be briefly reviewed here.

The first, with the longest tradition, has been aimed at including morality within formal logic. Hare [18] assumed moral sentences to belong to the general class of prescriptive languages, for which meaning come in two components: the *phrastic* which captures the state to be the case, or command to be made the case, and the *neustic* part, that determines the way the sentence is nodded by the speaker. While Hare did not provided technical details of his idea for prescriptive languages, in the same years Wright [31] developed deontic logic, the logical study of normative concepts in language, with the introduction of the monadic operators $O(\cdot)$, $F(\cdot)$, and $P(\cdot)$ for expressing obligation, prohibition and permission. It is well known that all the many attempts in this directions engender a set of logical and semantic problems, the most severe is the Frege-Geach embedding problem [12]. Since the semantics of moral sentences is determined by a non-truth-apt component, like Hare's neustic, it is unclear how they can be embedded into more complex propositions, for example conditionals. This issue is related with the elimination of the mental processes within the logic formalism, and in fact viable solutions are provided by proponents of expressivism, the theory that moral judgments express attitudes of approval or disapproval, attitudes that pertains to the mental world.

One of the best available attempt in this direction has been given by Blackburn [3] with variants of the deontic operators, like $H!(\cdot)$ and $B!(\cdot)$, that merely express attitudes regards their argument: "Hooray!" or "Boo!". Every expressive operator has its descriptive equivalent, given formally by the $|\cdot|$ operation. An alternative has been proposed by Gibbard [13] in possible worlds semantics, defining an equivalent expressivist friendly concept, that of *factual-normative world* $\langle W, N\rangle$ where W is an ordinary Kripke-Stalnaker possible world, while N, the system of norms, is characterized by a family of predicates like N-forbidden, N-required. If a moral sentence S is N-permitted in $\langle W, N\rangle$ then it is said to hold in that factual-normative world. Both proponents acknowledge the need of moving toward a mental inquire, but their aim did never translated into an effective attempt to embed genuine mental processes in a logic system.

The second account here sketched, was apparently motivated by filling the gap left by formal logic, the lack of the mental processes in morality. The idea that there exists a Universal Moral Grammar, that rules human moral judgments in analogy with Chomsky's Universal Grammar, was proposed several decades ago [26], but remained disregarded until recently, when resuscitated by Mikhail [22], who fleshed it out in great detail.

His fragment of Universal Moral Grammar is entirely fit to the "trolley dilemma", the famous mental experiment invented by Foot [10], involving the so-called doctrine of the double effect, which differentiates between harm caused as means and harm caused as a side effect, like deviating a trolley to save people, but killing another one. Mikhail refined importantly the trolley dilemma, by inventing twelve subcases that catch subtle differences. subjects. The model he

developed had the purpose of computing the same average responses given by subjects on the twelve trolley subcases. It is conceived in broad analogy with a grammatical parser, taking as input a structured description of the situation and a potential action, the moral grammar, and producing as output the decision if the potential action is permissible, forbidden, or obligatory. At the core of the grammar there is a "moral calculus", including rewriting rules from actions to moral effects.

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

3 Toward moral neurocomputing

It is manifest that for the internal enterprise, the modeling of choice should be neural computation, the attempt to imitate the computational process of the brain, in certain tasks. Neurocomputational approaches to morality were unfeasible without a coverage of empirical brain information [16]. A main realization to emerge from all the work done so far is that there is no unique moral module. There is no known brain region activated solely during moral thinking, while a relatively consistent set of brain areas that become engaged during moral reasoning, is also active in different non moral tasks. In brief, the areas involved in morality are also related to emotions, and decision making in general [15, 23, 6].

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 taking, are precious for any neurocomputational moral model. Reinforcement learning [27] is the reference formalization of the problem of how to learn from intermittent positive and negative events in order to improve action selection through time and experience. It has been the basis of early models using neuronlike elements [1], and the concepts of reinforcement learning have been later fitted into the biology of neuromodulation and decision making [8, 5].

The model GAGE [32] 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 [7], feelings that have become associated through experience with the predicted long-term outcomes of certain responses to a given situation. GAGE implementation of somatic-markers was based on Hebbian learning only, while reinforcement learning has

been adopted in ANDREA [21], 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 [19]. The overall architecture of these models have several similarities with those of [11], in which the orbitofrontal cortex interacts with the basal ganglia, but more oriented to dichotomic on/off decisions. A main drawback of all the models here mentioned is the lack of sensorial areas, that makes them unfit to be embedded even in the simplest form of environment in which a moral situation could be simulated.

4 The proposed model

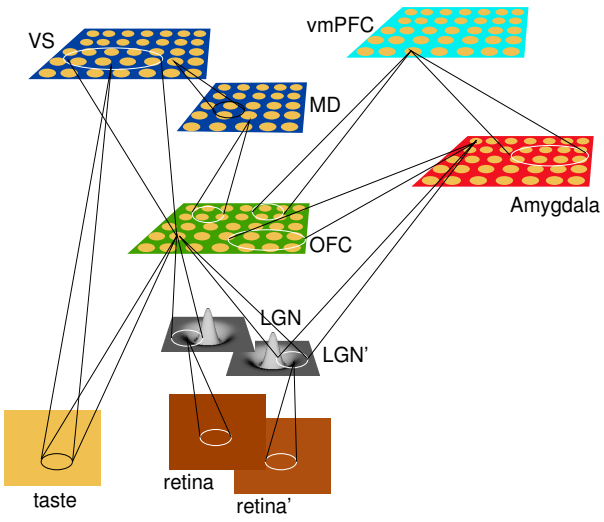

Fig. 1. Overall scheme of the model, composed by LGN (*Lateral Geniculate Nucleus*), V1 (*Primary Visual Area*), OFC (*OrbitoFrontal Cortex*), VS (*Ventral Striatum*), MD (*Medial Dorsal Nucleus*), Amyg (*Amygdala*), vmPFC (*ventromedial PreFrontal Cortex*).

The proposed model is able to simulate one specific moral situation, by including parts of the sensorial system, in connections to emotional and decision making areas. In the world seen by this artificial moral brain architecture there are three types of objects, two are neutral, and only one, resembling an apple, is edible, and its taste is pleasant. However, fruits in one quadrant of the scene are

forbidden, like belonging 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. The overall scheme 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 [2], and each cortical sheet adheres to the LISSOM (*Laterally Interconnected Synergetically Self-Organizing Map*) concept [30].

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 medial dorsal nucleus of the thalamus. 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.

4.1 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:

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

The vector fields $\boldsymbol{v}_i$, $\boldsymbol{e}_i$, $\boldsymbol{x}_i$ are circular areas of radius r_{A} for afferents, r_{E} for excitatory connections, r_{H} for inhibitory connections. The vector $\boldsymbol{a}_i$ is the receptive field of the unit i. Vectors $\boldsymbol{e}_i$ and $\boldsymbol{h}_i$ are composed by all connection strengths of the excitatory or inhibitory neurons projecting to i. The scalars γ_{A}, γ_{E}, γ_{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 L^1-norm.

4.2 Cortical components

The first circuit in the model learns the positive reward in eating fruits. The orbitofrontal cortex is the site of several high level functions, 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 [29], and others which respond to facial expressions [28], involved in social decision making [7].

OFC has forward and feedback connections with the Ventral Striatum, VS, which is the crucial center for various aspects of reward processes and motivation [17], and reprojects through MD, the medial dorsal nucleus of the thalamus, which, in turn, projects back to the prefrontal cortex. The global efficiency of the dopaminergic backprojections to OFC are modulated by a global parameter, used to simulate the hunger status of the model.

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 [7]. 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 [14]. It is involved in the development of morality, in a study [9] 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 [20]. In the model it is used specifically for capturing the negative emotion when seeing the angry face, a function well documented in the amygdala [4].

Fig. 2. Images seen by the model in the first phase of learning. On the left the patterns used for the development of the visual system. The other three images depict the objects that populate the simulated world: apples, +-shaped and ×-shaped.

4.3 First learning stages

The artificial brain is first exposed to a series of experiences, starting with a preliminary phase of development of the visual system with generic patterns, as those shown on the left in Fig. 2. These patterns mimic the retinal waves experienced before eye opening in humans, and allow the formation of retinotopy

and orientation domains in the model V1 area, similarly to the process described in [25]. When the visual system is mature, the model is presented with samples from the collection of three simple objects, in random positions, as shown in Fig. 2. At the same time their taste is perceived too, and only one of the objects, the apple, has a good taste. The connection loop between OFC, and the dopaminergic areas VS, MD, attain 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 this stage. The model will gradually become familiar with the objects, and learn how pleasant apples are, in its OFC model area. In order to characterize the ensemble activation pattern of the OFC neurons, and decode the objects categorization, a population code method is applied. The overall population is clustered according to those neurons, which were active in response to different classes of objects, compared to those which were not responsive, mathematical details are in [24].

Fig. 3. Neural coding of the three objects in the model OFC area: apple on the left, the ×-shaped in the center, and the ×-shaped on the right.

In Fig. 3 is shown the resulting coding of the three categories of objects in the OFC model area, with neurons that are selectively activated by objects of one class, independently on their position in space.

In a second stage the model receives additional experiences, that of the moral learning, with the same objects as stimuli. The model can choose between two possible behaviors: collect and eat an object, or refrain from doing it, a selection coded in the vmPFC component. Now, if the model decides to pick apples in a certain area of the world, that shown in the central image in Fig. 4, suddenly an angry face will appear, like those shown in the right of Fig. 4. Fruits in this portion of the space 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.

Now the amygdala gets inputs from both the OFC map and directly from the thalamus, when the angry face appears. There is an implicit reinforcement, with the negative reward embedded in the input projections to the amygdala.

Fig. 4. Images seen by the model in the second phase of learning. On the left an apple in a part of the world where it is allowed to pick it. The center image is an apple in the forbidden area, if the model attempts to pick it, the angry face shown on the right will suddenly appear.

4.4 Surviving without stealing

Finally, the developed artificial agent is embedded in its simple world, where all possible objects may randomly appear, and she can choose to grasp them or not. There is a parameter in the model which is used to modulate its state of hunger, in the dopaminergic circuit, which detailed equations are the following:

$$\begin{aligned} x^{(\mathrm{OFC})} = f\Big(& \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\mathrm{V1})} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\mathrm{V1})} + \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\circledcirc)} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\circledcirc)} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\circledcirc)} + \\ & \gamma_{\mathrm{A}}^{(\mathrm{OFC}\leftarrow\boxdot)} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{OFC}\leftarrow\boxdot)} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\boxdot)} + \gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} \boldsymbol{b}_{r_{\mathrm{B}}}^{(\mathrm{OFC})} \cdot \boldsymbol{v}_{r_{\mathrm{B}}}^{(\mathrm{MD})} + \\ & \gamma_{\mathrm{E}}^{(\mathrm{OFC})} \boldsymbol{e}_{r_{\mathrm{E}}}^{(\mathrm{OFC})} \cdot \boldsymbol{x}_{r_{\mathrm{E}}}^{(\mathrm{OFC})} - \gamma_{\mathrm{H}}^{(\mathrm{OFC})} \boldsymbol{h}_{r_{\mathrm{H}}}^{(\mathrm{OFC})} \cdot \boldsymbol{x}_{r_{\mathrm{H}}}^{(\mathrm{OFC})} \Big) \end{aligned} \tag{5}$$

$$\begin{aligned} x^{(\mathrm{VS})} = f\Big(& \gamma_{\mathrm{A}}^{(\mathrm{VS}\leftarrow\mathrm{OFC})} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{VS}\leftarrow\mathrm{OFC})} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\mathrm{OFC})} + \gamma_{\mathrm{A}}^{(\mathrm{VS}\leftarrow\boxdot)} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{VS}\leftarrow\boxdot)} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\boxdot)} + \\ & \gamma_{\mathrm{E}}^{(\mathrm{VS})} \boldsymbol{e}_{r_{\mathrm{E}}}^{(\mathrm{VS})} \cdot \boldsymbol{x}_{r_{\mathrm{E}}}^{(\mathrm{VS})} - \gamma_{\mathrm{H}}^{(\mathrm{VS})} \boldsymbol{h}_{r_{\mathrm{H}}}^{(\mathrm{VS})} \cdot \boldsymbol{x}_{r_{\mathrm{H}}}^{(\mathrm{VS})} \Big) \end{aligned} \tag{6}$$

$$x^{(\mathrm{MD})} = f\left(\gamma_{\mathrm{A}}^{(\mathrm{MD}\leftarrow\mathrm{VS})} \boldsymbol{a}_{r_{\mathrm{A}}}^{(\mathrm{MD}\leftarrow\mathrm{VS})} \cdot \boldsymbol{v}_{r_{\mathrm{A}}}^{(\mathrm{VS})} \right) \tag{7}$$

These two equations are just specialization of the general equation (1), for areas VS and MD. The afferent signals $\boldsymbol{v}^{(\mathrm{OFC})}$ come from the OFC model area, $\boldsymbol{v}^{(\boxdot)}$ is the taste signal, and [◎ the output of the LGN deferred in time, when a possibly angry face will appear. The output $x^{(\mathrm{MD})}$ computed in (7) will close the loop into the prefrontal cortex. The parameter $\gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})}$ 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.

A simulation is performed by letting the model meeting with random objects, at random positions in the world. Now there will be no more angry face in case the model steal an apple in the forbidden place, whoever, it is expected that the moral norm to avoid stealing will work, at least up to a certain level of hunger. There is no more learning in any area of the model. At every simulation step the modulation parameter is updated as following:

$$\gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} \leftarrow \begin{cases} \gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} - \chi & \text{when an apple is grasped} \\ \gamma_{\mathrm{B}}^{(\mathrm{OFC}\leftarrow\mathrm{MD})} + \phi & \text{otherwise} \end{cases} \tag{8}$$

Where χ is the amount of nutriment provided by an apple, and ϕ is the decrease of metabolic energy in time.

In Fig. 5 the decisions to grasp are shown, as a function of the hunger level, after 50000 simulation steps. Neutral objects are grasped occasionally, about one over three, almost independently from hunger. Allowed apples are grasped more frequently with hunger, every time with level over 0.1, while it can be seen the strong inhibition to grasp apples in the forbidden sector, with few attempts at extreme hunger level only, over 0.3.

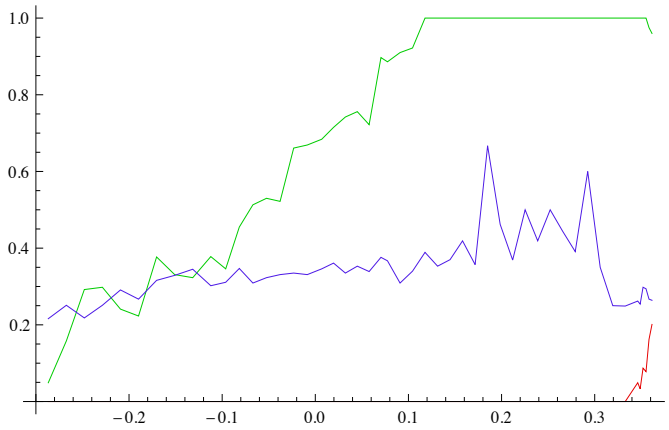

Fig. 5. Percentage of grasping decisions as a function of hunger level. Green: allowed apples, blue: neutral objects, red: forbidden apples.

In conclusion, we believe that the neurocomputational approach is an additional important path in pursuing a better understanding of morals, and this model, despite the limitation in its cortical architecture, and the crudely simplified external world, is a valid starting point. It picks up on one core aspect of morality: the emergence of a norm, not to steal, induced by a moral emotion. Obeying this norm is an imperative that supersedes other internal drives, like hunger, up to a certain extent. It has to be warned again, that morality is a collection of several, partially dissociated mechanisms, and the presented model is able to simulate only one kind of moral situation, the temptation of stealing food, and the potential consequent feelings of guilt. Further work will address other type of morality, that will need different scenarios to be simulated.

References

1. Barto, A., Sutton, R., Anderson, C.: Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man and Cybernetics 13, 834–846 (1983)

2. Bednar, J.A.: Topographica: Building and analyzing map-level simulations from Python, C/C++, MATLAB, NEST, or NEURON components. Frontiers in Neuroinformatics 3, 8 (2009)
3. Blackburn, S.: Spreading the Word. Oxford University Press, Oxford (UK) (1988)
4. Boll, S., Gamer, M., Kalisch, R., Büchel, C.: Processing of facial expressions and their significance for the observer in subregions of the human amygdala. NeuroImage 56, 299–306 (2011)
5. Bullock, D., Tan, C.O., John, Y.J.: Computational perspectives on forebrain microcircuits implicated in reinforcement learning, action selection, and cognitive control. Neural Networks 22, 757–765 (2009)
6. Casebeer, W.D., Churchland, P.S.: The neural mechanisms of moral cognition: A multiple–aspect approach to moral judgment and decision–making. Biology and Philosophy 18, 169–194 (2003)
7. Damasio, A.: Descartes' error: Emotion, reason and the human brain. Avon Books, New York (1994)
8. Dayan, P.: Connections between computational and neurobiological perspectives on decision making. Cognitive, Affective, & Behavioral Neuroscience 8, 429–453 (2008)
9. Decety, J., Michalska, K.J., Kinzler, K.D.: The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. Cerebral Cortex 22, 209–220 (2012)
10. Foot, P.: The problem of abortion and the doctrine of the double effect. Oxford Review 5, 5–15 (1967)
11. Frank, M.J., Scheres, A., Sherman, S.J.: Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. Philosophical transactions of the Royal Society B 362, 1641–1654 (2007)
12. Geach, P.T.: Assertion. The Philosophical Review 74, 449–465 (1965)
13. Gibbard, A.: Wise Choices, Apt Feelings – a theory of normative judgment. Harvard University Press, Cambridge (MA) (1990)
14. Gläscher, J., Hampton, A.N., O'Doherty, J.P.: Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cerebral Cortex 19, 483–495 (2009)
15. Greene, J.D., Haidt, J.: How (and where) does moral judgment work? Trends in Cognitive Sciences 6, 517–523 (2002)
16. Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D.: fMRI investigation of emotional engagement in moral judgment. Science 293, 2105–2108 (2001)
17. Haber, S.N.: Neural circuits of reward and decision making: Integrative networks across corticobasal ganglia loops. In: Mars, R.B., Sallet, J., Rushworth, M.F.S., Yeung, N. (eds.) Neural Basis of Motivational and Cognitive Control, pp. 22–35. MIT Press, Cambridge (MA) (2011)
18. Hare, R.M.: The Language of Morals. Oxford University Press, Oxford (UK) (1952)
19. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decisions under risk. Econometrica 47, 313–327 (1979)
20. LeDoux, J.E.: Emotion circuits in the brain. Annual Review of Neuroscience 23, 155–184 (2000)
21. Litt, A., Eliasmith, C., Thagard, P.: Neural affective decision theory: Choices, brains, and emotions. Cognitive Systems Research 9, 252–273 (2008)
22. Mikhail, J.: 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 (2009)

23. Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J.: The neural basis of human moral cognition. Nature Reviews Neuroscience 6, 799–809 (2005)
24. Plebe, A.: 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 (2014)
25. Plebe, A., Domenella, R.G.: Object recognition by artificial cortical maps. Neural Networks 20, 763–780 (2007)
26. Rawls, J.: A Theory of Justice. Belknap Press of Harvard University Press, Cambridge (MA) (1971)
27. Rescorla, R.A., Wagner, A.R.: A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (eds.) Classical Conditioning II: Current theory and research, pp. 64–99. Appleton Century Crofts, New York (1972)
28. Rolls, E., Critchley, H., Browning, A.S., Inoue, K.: Face-selective and auditory neurons in the primate orbitofrontal cortex. Experimental Brain Research 170, 74–87 (2006)
29. Rolls, E., Critchley, H., Mason, R., Wakeman, E.A.: Orbitofrontal cortex neurons: Role in olfactory and visual association learning. Journal of Neurophysiology 75, 1970–1981 (1996)
30. Sirosh, J., Miikkulainen, R.: Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. Neural Computation 9, 577–594 (1997)
31. Von Wright, G.H.: Deontic logic. Mind 60, 1–15 (1951)
32. Wagar, B.M., Thagard, P.: Spiking Phineas Gage: A neurocomputational theory of cognitiveaffective integration in decision making. Psychological Review 111, 67–79 (2004)