CHAPTER

23

Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans

S. Palminteri[1,2], *M. Pessiglione*[3,4]

[1]University College London, London, United Kingdom; [2]Ecole Normale Supérieure, Paris, France; [3]Institut du Cerveau et de la Moelle (ICM), Inserm U1127, Paris, France; [4]Université Pierre et Marie Curie (UPMC-Paris 6), Paris, France

Abstract

Approaching rewards and avoiding punishments are core principles that govern the adaptation of behavior to the environment. The machine learning literature has proposed formal algorithms to account for how agents adapt their decisions to optimize outcomes. In principle, these reinforcement learning models could be equally applied to positive and negative outcomes, ie, rewards and punishments. Yet many neuroscience studies have suggested that reward and punishment learning might be underpinned by distinct brain systems. Reward learning has been shown to recruit midbrain dopaminergic nuclei and ventral prefrontostriatal circuits. The picture is less clear regarding the existence and anatomy of an opponent system: several hypotheses have been formulated for the neural implementation of punishment learning. In this chapter, we review the evidence for and against each hypothesis, focusing on human studies that compare the effects of neural perturbation, following drug administration and/or pathological conditions, on reward and punishment learning.

Good and evil, reward and punishment, are the only motives to a rational creature: these are the spur and reins whereby all mankind are set on work, and guided.

These famous words by John Locke suggest that rewards and punishments are not on a continuum from positive to negative: they pertain to distinct categories of events that we can imagine or experience. Indeed rewards and punishments trigger different kinds of subjective feelings (such as pleasure versus pain or desire versus dread) and elicit different types of behaviors (approach versus avoidance or invigoration versus inhibition). These considerations might suggest the idea that rewards and punishments are processed by different parts of the brain. In this chapter we examine this idea in the context of reinforcement learning, a computational process that could in principle apply equally to rewards and punishments. We start by summarizing the computational principles underlying reinforcement learning (Box 23.1 and Fig. 23.1) and by describing typical tasks that implement a comparison between reward and punishment learning (Box 23.2 and Fig. 23.2). Then we expose the current hypotheses about the possible implementation of reward and punishment learning systems in the brain (Fig. 23.3). Last,

BOX 23.1

COMPUTATIONAL MODELS OF REINFORCEMENT LEARNING

The first reinforcement learning (RL) models come from the behaviorist tradition, in the form of mathematical laws describing learning curves [82] or formal descriptions of associative conditioning [2]. Subsequently, in the 1980s, computational investigation of RL received a significant boost when it grabbed the attention of machine learning scholars, who were aiming at developing algorithms for goal-oriented artificial agents [1]. In the

Continued

Decision Neuroscience
http://dx.doi.org/10.1016/B978-0-12-805308-9.00023-3

Copyright © 2017 Elsevier Inc. All rights reserved.

BOX 23.1 (*cont'd*)

machine learning literature, the typical RL problem involves an agent navigating through a series of states (s) by performing actions (a), while collecting some numeric reward (r). The goal of the agent is to select actions that maximize the future cumulative reward (also referred to as "return"). The typical RL algorithm has therefore two main functions: the value function, which stores reward predictions, and the policy function, which determines which action has to be taken to maximize the reward.

A variety of solutions to the RL problem have been proposed. The most relevant RL models for psychologists and neurobiologists revolve around the notion of reward-prediction error, which is equivalent to temporal difference error in most choice tasks, in which there is only one transition step between stimuli and outcomes. Reward-prediction error (RPE) is the difference between obtained and expected outcomes. Thus, after receiving a reward r at trial t, an RPE δ is calculated based on the current estimate of the state value $V(s)$ as follows:

$$\delta_t = r_t - V(s)_t \tag{23.1}$$

This error term is subsequently used to update (improve) the reward prediction through a learning rule. The most commonly used is the delta rule, in which the impact of each RPE on future expectation is scaled by a learning rate α:

$$V(s)_{t+1} = V(s)_t + \alpha^* \delta_t \tag{23.2}$$

In RL models action selection can rely on "direct" or "indirect" policy functions. Direct policy implies that, instead of representing only state-based values [$V(s)$], the agent represents values that are both action- and state-dependent [$Q(s,a)$]. Whereas state value represents the reward expected in a given situation, action values represent the reward expected from taking a particular action in a given situation. Action values can also be learned via prediction errors and delta rules, and directly compared to make a choice, as implemented in the Q-learning model—a model very frequently used in human and animal studies [83] (Fig. 23.1A, left). Indirect policy involves two separate representations for value prediction and action propensity, as famously implemented in the actor–critic model (Fig. 23.1A, right). In this model, the actor makes choices by comparing state-dependent action propensities [$\pi(s,a)$]. The obtained reward r generates an RPE δ, relative to the state value stored by the critic [$V(s)$]. The RPE is then used to improve ("criticize") future reward expectations, as in Eq. (23.2), as well as action propensities in the following equation:

$$\pi(s,a)_{t+1} = \pi(s,a)_t + \alpha^* \delta_t \tag{23.3}$$

An important problem that all RL algorithms must address is the trade-off between exploiting current knowledge and exploring alternative options. This exploitation/exploration trade-off is particularly relevant in probabilistic and changing environments, in which sticking to first impressions may prove ruinous. The simplest way to address this issue is to allow some stochasticity in the decision process. Thus, instead of systematically picking the highest value action (hard maximization or greedy decision rule), a softmax decision rule has been proposed [84] in which the probability of choosing A over B is a sigmoid function of the value difference between A and B:

$$P(A) = 1/(1+\exp((Q(B)-Q(A))/\beta)) \tag{23.4}$$

Here β is a temperature parameter that adjusts the steepness of the sigmoid function. The softmax function implies that exploration is maximal when $Q(A) \approx Q(B)$. Note that this is the same function that is used in logistic regression, in which the β weight on the value difference would be equivalent to an inverse temperature.

Computational models are useful for the experimental exploration of human RL abilities in many respects. First, they may be used to generate trial-by-trial estimates of value prediction and prediction errors, which can then be mapped to brain activity using fMRI, an approach termed model-based fMRI [85]. Second, computational models may help finesse the analysis of learning deficits induced by drugs and lesions, compared to aggregate behavioral measures (Fig. 23.1B). For example, consider a case in which two different treatments are shown to impair instrumental learning, as evidenced by a decrease in correct choice rate compared to placebo. We might be tempted to conclude that the two drugs have "similar effects." However, computational analysis may reveal that one drug affects the learning rate and the other the choice temperature parameter, thus dissociating their effects on the update versus selection process [86].

we review the evidence in favor or against the various hypotheses, focusing on studies in humans that compare reward and punishment learning and that employ approaches that assess causal implications, by observing the behavioral effects of pharmacological manipulations and brain lesions.

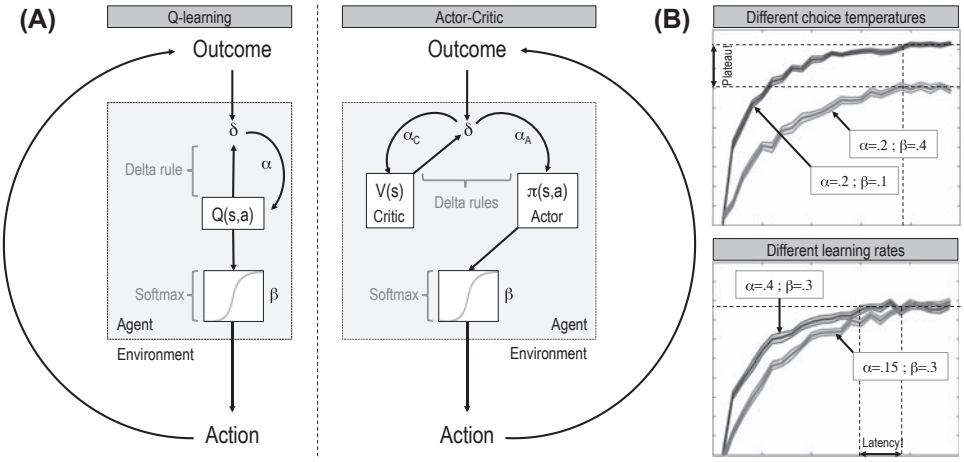

FIGURE 23.1 **Basic computational models of reinforcement learning.** (A) Computational architectures of standard models using direct and indirect policy rules (*Q*-learning and actor–critic models, respectively). In both architectures the decision process (policy) is modeled with a softmax function, whose exploration/exploitation trade-off is governed by parameter β (temperature). In the *Q*-learning model decisions are made by comparing *Q*-values, which are action-specific estimates of the expected future reward. When the outcome is received, a prediction error δ is calculated as the difference between the chosen action *Q*-value and the actual reward. The prediction error is then used to update the chosen action *Q*-value via a delta rule adjusted by parameter α (learning rate). In the actor–critic model decisions are made by comparing policy values π, which are action-specific estimates of action probabilities, stored by the actor component. When the outcome is received, a prediction error δ is calculated as the difference between the state value, stored by the critic component, and the actual reward. The prediction error is then used in a delta rule to update reward prediction (ie, the state value) as well as the action probability (ie, the π-value), via different learning rates. (B) Macroscopic effects of varying key *Q*-learning parameters (learning rate α and choice temperature β) on average learning curves. Crucially, the learning rate affects only the latency, ie, the number of trials required to reach a given performance level, whereas the temperature also affects the plateau, ie, the performance level after convergence.

BOX 23.2

BEHAVIORAL TASKS USED TO COMPARE REWARD AND PUNISHMENT LEARNING

The aim of such tasks is to dissociate valence-specific and valence-independent processes. It is important to implement the comparison within the same task to avoid confounds with details of the design and to avoid framing effects. Indeed subjects might reframe their expectations if they realize they are in a reward- or punishment-learning task, ie, they might change their reference point and, for instance, take an absence of reward as a punishment or an absence of punishment as a reward [76–78,87]. Contrasting reward and punishment in the same protocol entails the challenging problem of comparing stimuli that do not necessary share the same properties and whose values are not necessarily in the same range. A simple solution is to opt for secondary reinforcers such as money or even abstract "points." In this case, rewards and punishments share the same sensory properties and, being numeric in nature, they can be directly fed to model-based analyses. Yet the generalizability to other forms of reinforcers, perhaps more natural for the brain, of results obtained using money or points is not automatically granted.

In the following we focus on instrumental (or operant) learning tasks but note that Pavlovian (or classical) learning tasks could also be used. In Pavlovian tasks, the occurrence of reinforcers is not contingent on the behavior of the subject, who can remain entirely passive. In the reinforcement learning (RL) framework, Pavlovian tasks elicit only a state value function, which can be used to fit implicit measures of learning such as pupil diameter or skin conductance response. Yet these measures are noisy, and model fitting implies specifying a function that relates them to state values, which is not as straightforward as in the case of choices [8]. Also, for Pavlovian tasks that do not require any overt behavior, it may be harder to control the engagement of subjects, an issue that is particularly problematic with patients. However, Pavlovian tasks avoid the issue of possible confounds with motor responses, which must be carefully orthogonalized with respect to outcome valence in instrumental tasks. This is because reward obtainment and punishment avoidance might be more naturally associated with "go" and "no-go" responses, respectively [88,89].

Instrumental learning tasks in humans have most frequently taken the form of two-armed bandits. Subjects are repeatedly presented with a choice between two abstract stimuli (often fractal images or letters taken from exotic alphabets) representing the two available actions.

Continued

BOX 23.2 (*cont'd*)

Their task is to find out, by trial and error, the action with the highest expected value. One influential design (the Hiragana task, Fig. 23.2A) differentiates between a training and a test phase [38,45]. During the training phase subjects are asked to play several two-armed bandits (often three) until they reach a performance criterion. Crucially, the average value (ie, the policy-independent state value) of the bandits is zero because the two stimuli have reciprocal (and symmetrical) probabilities of winning/losing a point. In other terms, choices observed during the training phase do not discriminate between reward and punishment learning, because both outcomes can be experienced with the same bandit. After training, subjects are asked to recognize the most advantageous stimulus among new pairings of stimuli, in the absence of feedback. Choices in this phase are taken as indicators of learning that occurred during the training phase. The capacity to select the best possible stimulus is considered a measure of reward learning and the capacity to avoid the worst stimulus a measure of punishment learning.

Another influential design (the Agathodaimon task, Fig. 23.2B) directly considers instrumental choices made during learning, instead of postlearning preferences [25,66]. In this task, subjects are also asked to play several two-armed bandits (at least two, sometimes four). The crucial difference relative to the other task is that winning and losing are now opposed to neutral outcomes. So the average value of the bandits is different from zero, being either positive (in the reward maximization conditions) or negative (in the punishment minimization conditions). In other terms, this task directly discriminates between reward-seeking and punishment-avoidance performance, because both outcomes cannot be experienced in the same bandit. The correct response rates in the two conditions are thus considered as measures of reward and punishment learning, respectively. Importantly, this task allows fitting RL models on choice learning curves separately for reward and punishment conditions, and therefore characterizing valence-specific deficits in computational terms. Also, this task allows disentangling between two oppositions that were confounded in the previous task: reward versus punishment (outcome valence) and positive versus negative prediction errors (which are both present in the two conditions).

THE NEURAL CANDIDATES FOR REWARD AND PUNISHMENT LEARNING SYSTEMS

Reinforcement learning (RL) refers to a class of processes through which an agent builds associations between stimuli and actions under the influence of rewards or punishments, ie, events that possess a positive or negative value for the agent's well-being. RL processes have been formally captured by a variety of algorithms in the machine learning literature (see Box 23.1 and Fig. 23.1). The typical RL algorithm learns, by trial and error, to select between available actions in an environment characterized by a given set of possible states, so as to maximize some notion of cumulative reward [1]. A key principle, inherited from the animal learning literature, that is common to most RL algorithms is learning being driven by reward-prediction error (RPE), a sort of signed surprise defined as the difference between expected and obtained rewards [2].

Electrophysiology studies have consistently reported RPE encoding in the dopaminergic midbrain areas of both human and nonhuman primates [3,4]. More precisely, it has been reported that the firing rate of dopamine (DA) neurons positively and parametrically scales with RPE [5]. Neuroimaging studies in humans have confirmed and extended these results by showing functional magnetic resonance imaging (fMRI) correlates of RPE in the midbrain [6], as well as in the main DA projection sites, such as the striatum and the prefrontal cortex, especially in their ventral parts [7–9].

In RL algorithms, there is no reason why the reward term could not take negative values, and hence capture the notion of punishment. This implies that, at least in principle, the same computations could be used for reward and punishment learning. However, physiological constraints suggest that a single neural system might not be able to encode in an equally efficient manner both RPEs and punishment-prediction errors (PPEs). This is simply because firing rate cannot be negative, and therefore neurons that have low spontaneous activity, as those in the midbrain, do not have sufficient range to encode PPE precisely with decreasing firing rate [10]. To obviate this physiological constraint, one solution is to assume that an opponent system might positively respond to PPE, just as the dopaminergic midbrain does for RPE [11]. So far,

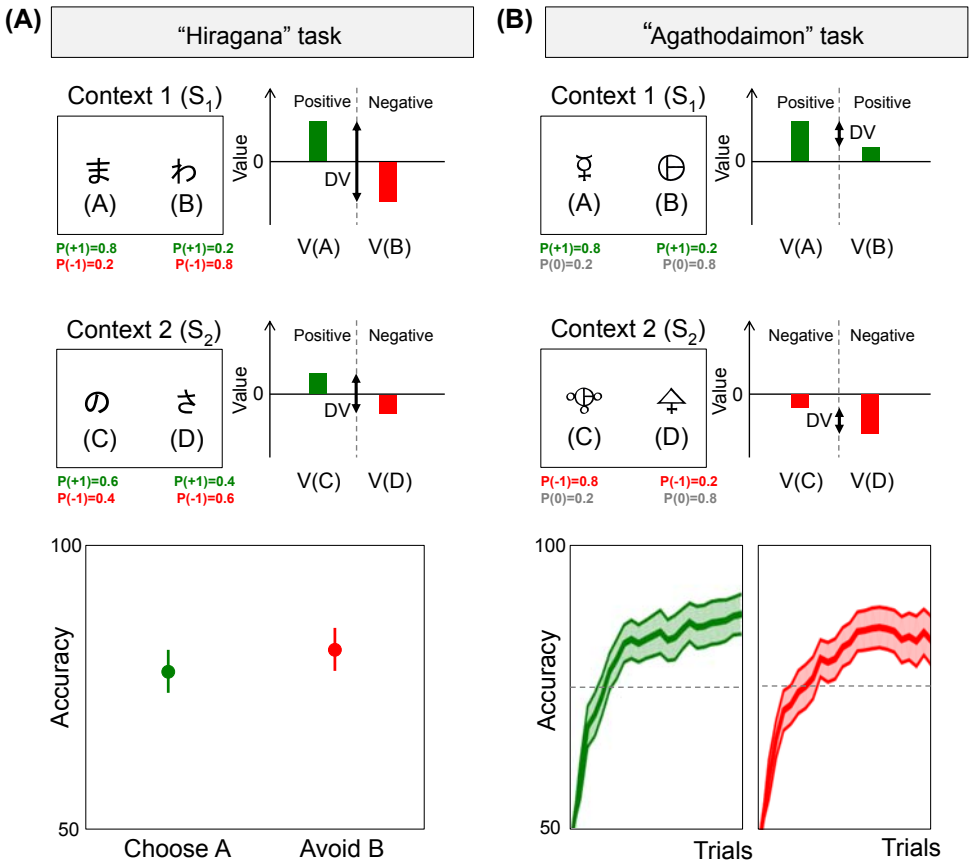

FIGURE 23.2 **Two classical tasks used to compare reward and punishment learning.** Presented for both tasks are decision screens in two possible contexts (pairs of symbols), the probabilistic contingencies associated with each symbol, the two option values (DV is the decision value, ie, the difference between the two options), and the main performance measure (choice accuracy) expected from a normal subject. Note that values are the actual values that subjects have to learn, before learning they are equal to zero. (A) The Hiragana task is designed so as to reveal in a test session the type of learning (reward seeking versus punishment avoidance) that was operant during the training session [38]. During the training session subjects are presented with fixed pairs of options (typically three pairs), materialized by Hiragana symbols and associated with different, reciprocal probabilities of winning or losing, P(+1) and P(−1). During the test session subjects are asked to identify the best option, among novel binary combinations, in the absence of feedback. The capacity to correctly identify the best option (choose A) and reject the worst (avoid B) is taken as a measure of the capacity to learn from positive and negative prediction errors. (B) The Agathodaimon task is designed to compare reward and punishment learning directly during the training session [25]. Subjects are also presented with fixed pairs of symbols (typically two pairs), now materialized by Agathodaimon symbols, with the crucial difference that rewards and punishments are never mixed within a pair. Some pairs of options are associated with reciprocal probabilities of winning or getting nothing, ie, P(+1) and P(0), and others with reciprocal probabilities of losing or getting nothing, ie, P(−1) and P(0). Typically, the rate of correct choice (ie, choosing the most rewarding or the least punishing option) is extracted on a trial-by-trial basis to assess the capacity to learn from rewards versus punishments.

several hypotheses have been formulated concerning the neural implementation of this tentative opponent system, but this remains extremely controversial [12–15] (see also Ref. [16] of the present volume). We have identified and describe here four main hypotheses (see also Fig. 23.3).

Hypothesis 1: No Opponent System

A first hypothesis is that there is actually no opponent system and that punishment avoidance is also resolved by the midbrain DA system within the basal ganglia circuits. It has been argued that, whereas phasic burst in DA activity encodes positive prediction errors, the duration of pauses in DA activity might encode negative prediction errors [17]. In this framework, an important role may be played by the habenula, an epithalamic nucleus whose activity has been shown to provide inhibitory input to the midbrain DA neurons following reward omission in monkeys [18]. Consistently, the habenula has been shown, in humans, to encode aversive events such as electric shocks and to have an impact on striatal activity [19].

Hypothesis 2: Dopaminergic Opponent System

A second hypothesis also supposes that avoidance learning is driven by DA activity, but thanks to a

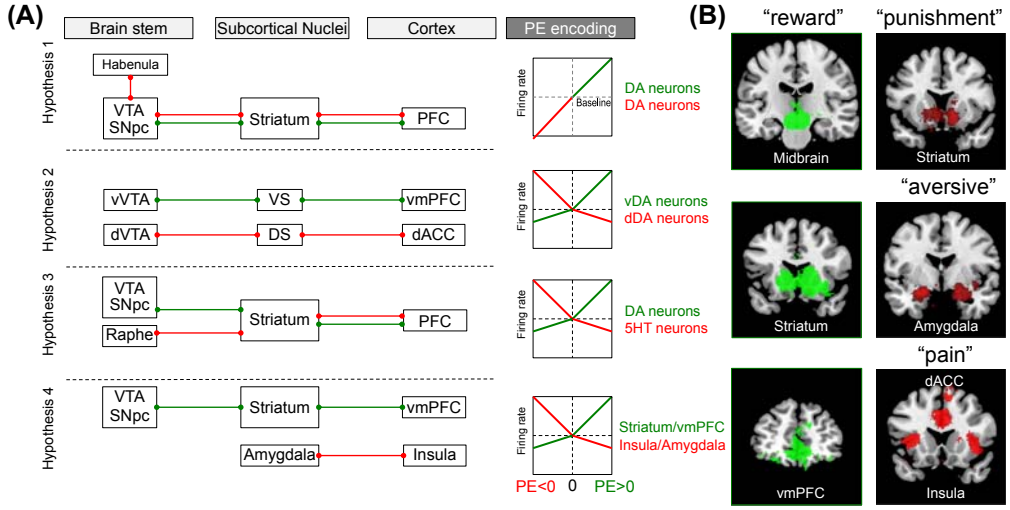

FIGURE 23.3 **Neural implementation of the opponency between reward and punishment learning.** (A) Various hypotheses concerning the neural implementation of punishment avoidance (in red), as opposed to reward seeking (in green). For each hypothesis, the key regions and connections of each opponent system are shown on the left, with their theoretical pattern of activity as a function of prediction error (PE) plotted on the right. (B) Maps resulting from automatized large-scale meta-analyses as implemented in Neurosynth [90]. Meta-analyses confirm the implication of the VTA–VS–vmPFC circuit in reward encoding, as is assumed in most hypotheses. Negative affects such as "punishment," "aversive," and "pain" involve both similar (striatum) and specific brain regions (notably dACC, insula, amygdala). Note that Neurosynth treats similarly positive and negative correlations between experimental factors and brain activity, leaving undecided whether the striatal representation of "punishment" is driven by activation with avoided punishment (hypothesis 1) or received punishment (hypothesis 2). *5HT*, serotonin; *DA*, dopamine; *dACC*, dorsal anterior cingulate cortex; *dDA*, dorsal dopamine; *DS*, dorsal striatum; *dVTA*, dorsal VTA; *PFC*, prefrontal cortex; *SNpc*, substantia nigra pars compacta; *vDA*, ventral dopamine; *vmPFC*, ventromedial PFC; *VS*, ventral striatum; *VTA*, ventral tegmental area; *vVTA*, ventral VTA.

different subset of midbrain neurons that positively encode punishments [20]. This hypothesis is based on electrophysiological observations that an anatomically segregated population of DA neurons in the nonhuman primate midbrain positively respond to aversive stimuli [21]. Consistently, human fMRI studies have shown positive encoding of punishment anticipation and PPE in the ventral tegmental area (VTA) and downstream in the striatum during aversive conditioning tasks [22–24]. fMRI data are generally consistent with the idea of a functional gradient, such that the ventral parts of the frontostriatal circuits would be preferentially concerned with reward seeking and dorsal parts with punishment avoidance [22,25,26].

Hypothesis 3: Serotoninergic Opponent System

A third hypothesis states that the neuromodulator serotonin (5-HT) could play the role of an opponent signal by encoding PPE [11]. A vast body of literature in rodents, linking the serotoninergic system (especially the dorsal raphe) to behavioral inhibition and "fight or flight" responses (generally induced by aversive events), originally motivated this hypothesis [27,28]. Further supporting this idea, at the electrophysiological level, 5-HT has been shown to antagonize DA function in the VTA and striatum [29,30].

Hypothesis 4: Other Opponent Systems

Finally, the fourth hypothesis postulates that punishment avoidance involves circuits outside the frontostriatal projections of the brain-stem neuromodulator systems. According to this hypothesis (which would be more appropriately considered as a collection of hypotheses), punishment learning is mediated by aversive signals encoded in other cortical and subcortical areas, such as the insula and amygdala (see also Ref. [31] in the present volume). A consistent body of electrophysiological, pharmacological, and lesion studies in animals supports the implication of these regions in punishment avoidance [32,33]. These results from animal studies align with some fMRI studies in humans, as well as with meta-analyses [34–37].

EVIDENCE FROM DRUG AND LESION STUDIES

For the sake of simplicity, we have assumed that the opponent systems encoding rewards and punishments both have a better precision (or gain) with increasing firing rates. This makes the prediction that damage to a subset of this system should preferentially degrade either reward or punishment learning and therefore produce effects on choice behavior that should interact with

outcome valence. In the following we examine this prediction, under the four hypotheses regarding the implementation of the opponent system underlying punishment learning. We focus on tasks that were employed in humans to compare reward and punishment learning directly (see examples in Box 23.2 and Fig. 23.2).

The first hypothesis states that midbrain DA bursts and dips are necessary and sufficient for reward and punishment learning, respectively. Thus, according to this hypothesis, enhancing DA function should increase reward learning to the detriment of punishment learning and DA blocking should produce the opposite effects. These predictions have been verified in Parkinson's disease (PD) patients [38]. PD is characterized by DA neuronal loss and treated with DA enhancers, either metabolic precursors of DA or direct agonists of DA receptors [39]. A group of patients was tested twice (once ON and once OFF medication) with a probabilistic learning task (the Hiragana task in Fig. 23.2A). Results showed a significant medication by valence interaction, with ON-PD patients being better at reward learning than OFF-PD patients, and vice versa for punishment learning. This result has been interpreted within the framework of a neural network model of the basal ganglia that formalizes action selection as the result of a competition between a direct "go" and an indirect "no-go" pathway [40]. On one side, the go pathway expresses D1 DA receptors and is reinforced by DA bursts, leading to an increased probability of choosing options followed by reward. On the other side, the no-go pathway expresses D2 DA receptors and is reinforced by DA dips, leading to a reduced probability of choosing options followed by punishment. The interaction between medication status and outcome valence has been replicated several times in PD patients [41,42]. Another study further suggested that improvement in reward seeking observed under pro-DA modulation is specific to PD patients with DA dysregulation syndrome, whereas impairment of punishment avoidance stands only for nondysregulated patients [43]. This neurobiological model has received support from genetic studies, indicating specific roles for D1 and D2 polymorphisms in reward seeking and punishment avoidance behaviors, respectively [44,45].

Thus, investigations of RL abilities in PD with and without dopaminergic treatment were consistent with the idea that DA dips are necessary for punishment learning. The questions remained (1) whether these effects arose from the modulation of explicit learning processes or rather from implicit processes and (2) whether these effects were specific to PD patients and their medication or can be generalized to other conditions and treatments. To examine these questions, we adapted an instrumental learning task (the Agathodaimon task in Fig. 23.2B) such that symbolic cues indicating the state of the environment were not consciously perceived, and we tested patients with Tourette syndrome (TS) in addition to PD patients [46,47]. TS is an interesting model for studying RL because it is characterized by hyper-DA symptoms and treated with neuroleptics, an anti-DA medication (Fig. 23.4A) [48]. Our results concerning PD replicated previous findings, indicating that the interaction between medication status and outcome valence holds for implicit learning processes (Fig. 23.4B). Interestingly, TS patients displayed the opposite double dissociation, with OFF-TS patients being better at reward seeking and ON-TS patients at punishment learning. Thus, untreated PD and treated TS might receive the same interpretation: DA levels are too low for RPE-related DA bursts to reinforce approach behavior (Fig. 23.4C). Reciprocally, in treated PD and untreated TS, DA levels might be too high for PPE-related DA dips to reinforce avoidance behavior.

However, despite this suggestive evidence for the implication of DA dips in punishment learning, other studies directly challenged these findings. In fact, while enhancing reward learning by increasing DA levels has been almost systematically observed, results regarding punishment learning have been less consistent, with several studies showing no effect of dopaminergic drugs on avoidance behavior, even with doses that were efficient on reward seeking [25,49–51]. Some of these studies fitted RL models to the observed choices to identify the computational parameter that would best capture drug effects. Interestingly, the positive effect of levodopa on reward learning was best accounted for by increasing the reward parameter, and not the learning rate that modulates RPE [25]. Thus, it could be that dopaminergic drugs just amplify reward representation, without affecting learning per se. By contrast, there was no significant effect on the punishment parameter.

To our knowledge, there is no pharmacological evidence that pro-DA drugs could improve punishment avoidance, as would be predicted by the second hypothesis, according to which a distinct (dorsal) population of DA neurons positively encodes aversive signals and underpins punishment learning. A corollary of this second hypothesis is the idea of a selective implication of the anterior caudate (commonly referred to as "dorsal striatum" in human fMRI literature) in punishment processing [22], whereas the nucleus accumbens (commonly referred to as "ventral striatum" in human fMRI literature) would be more involved in reward processing. To test this idea, we administered a probabilistic instrumental learning task (Agathodaimon task, Fig. 23.2B) to patients with Huntington disease (HD). This disease is a rare genetic condition characterized by choreic movements and caused by degeneration of

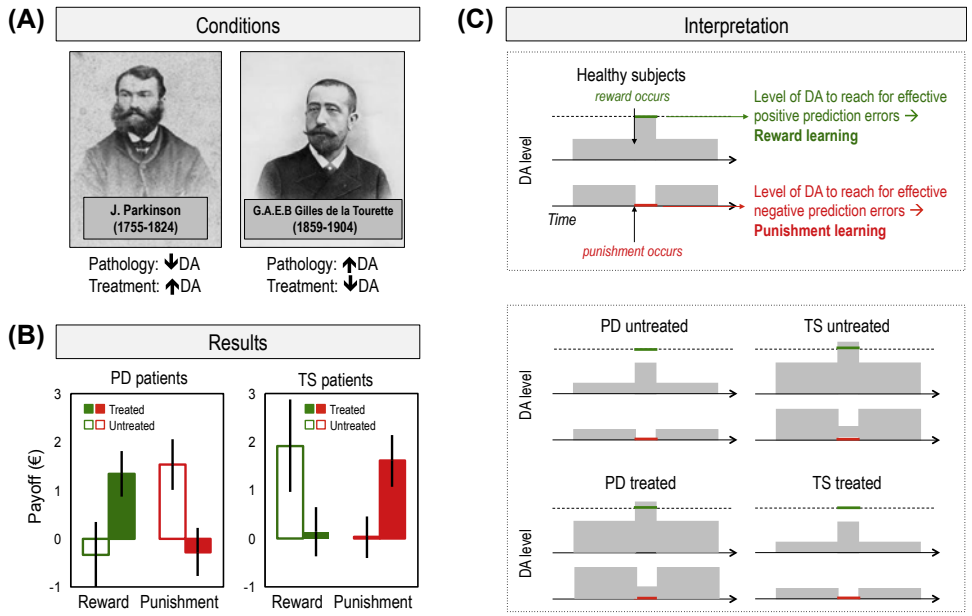

FIGURE 23.4 **Double dissociation of dopaminergic medication effects on reward and punishment learning.** (A) Parkinson's disease *(PD)* and Tourette syndrome *(TS)*, named after James Parkinson and Georges Albert Édouard Brutus Gilles de la Tourette, have proven to be useful models to study the role of dopamine (DA) in reinforcement learning. They represent opposite conditions in terms of symptoms (hypo- vs hyperkinesia), pathophysiology (hypo- vs hyperdopaminergia), and medication (pro- vs antidopaminergic) [91]. (B) Histograms in each graph show additional correct choices (in euros) in the reward-seeking (green) and punishment-avoidance (red) conditions relative to a neutral condition, observed in the Agathodaimon task. The results show an interaction between outcome valence (positive or negative) and medication status (treated or untreated) in both conditions, with opposite patterns [47]. (C) A schematic interpretation of DA activity (gray) following positive and negative outcomes, in healthy subjects as well as PD and TS patients. The *green and red lines*, respectively, represent the level to reach (either above or below the baseline) to express a signal strong enough to induce either positive or negative value update. The pathological conditions and pharmacological manipulations shift the baseline, either downward (untreated PD and treated TS), such that positive prediction errors fail to induce reward learning, or upward (treated PD and untreated TS), such that negative prediction errors fail to induce punishment learning [92]. This schematic is based on the results originally reported by Schultz and colleagues (1997) and largely inspired by the theory of Frank and colleagues (2004).

the striatum (Fig. 23.5A) [52]. The neural degeneration starts in the dorsal parts of the striatum, in the caudate head mostly, before the motor symptoms become apparent [53]. This makes presymptomatic HD a relevant lesion model to investigate dorsal striatal function. Our results were consistent with the idea that the dorsal striatum is specifically involved in punishment learning (Fig. 23.5B). However, our computational analyses revealed that the deficit observed in presymptomatic HD patients was best explained by increasing choice stochasticity, and not reducing the punishment parameter or punishment learning rate. Thus the dorsal striatum (anterior caudate) system might not be implicated in learning per se but in selecting between actions that lead to negative outcomes (the lesser of two evils). This could relate to the notion that dorsal prefrontostriatal circuits are responsible for response inhibition or avoidance behavior.

Despite its great theoretical appeal, the third hypothesis, which assigns to 5-HT the role of an opponent neuromodulator system, has received mixed evidence in human studies. While some studies did provide support for a specific role of 5-HT in punishment learning, other studies found nonspecific effects or even provided evidence for a specific role in reward learning [54–59]. Such inconsistent results have called for a revision of the original theory, which now incorporates a behavioral dimension—approach versus withdrawal—in addition to that of outcome valence—reward versus punishment [12,60]. In this theoretical framework, 5-HT would be needed to avoid punishment through response inhibition (no-go), but not when avoiding punishment implies response invigoration (go). Yet this addendum does not resolve every discrepancy reported in the literature. For instance, using the same subliminal learning task as in PD and TS patients [47], we found that 5-HT reuptake inhibitors, given as treatment for obsessive compulsive disorder, improved reward and punishment learning to similar extents and whether the response was implemented as a go or a no-go [58]. The complexity of the role of 5-HT in RL is further highlighted by its proven implication in the temporal discounting of reward [61,62]. It has even been argued that because the serotoninergic system is much more anatomically widespread

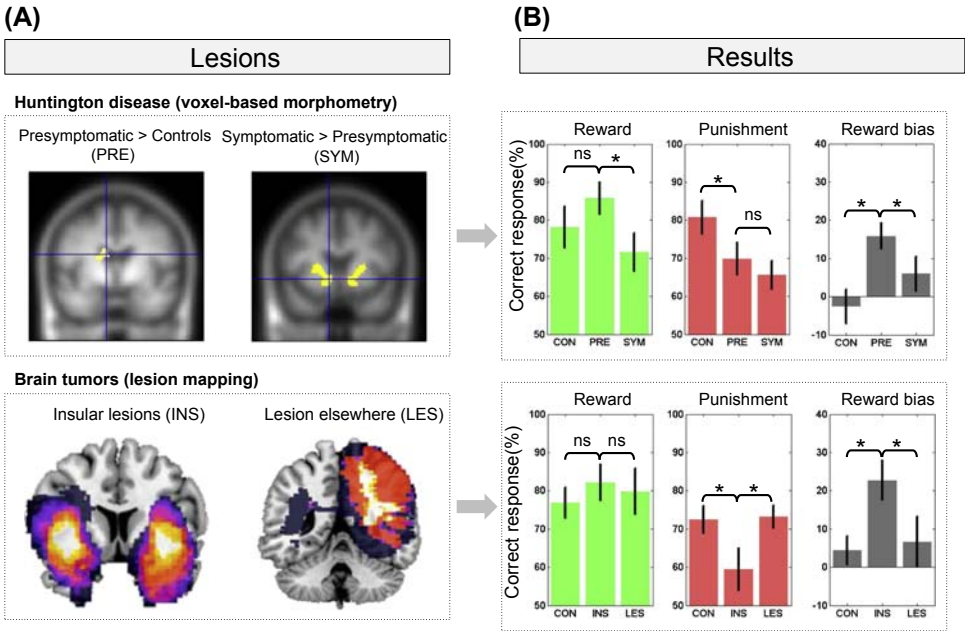

FIGURE 23.5 **The causal role of the anterior insula and dorsal striatum in punishment (but not reward) learning.** (A) Mapping of brain lesions. Striatal lesions were caused by neurodegenerative processes occurring in presymptomatic (*PRE*; dorsal striatum) and early symptomatic (*SYM*; both dorsal and ventral striatum) patients with Huntington disease. Insular lesions (INS) as well as control lesions (LES) outside the insula were due to low-grade gliomas. (B) Behavioral results from the Agathodaimon task. Both the PRE and the INS groups showed reduced punishment learning and consequently a positive reward bias (correct choice rate in reward minus punishment context). This demonstrates the critical and specific implication of both dorsal striatum and anterior insula in punishment learning.

and genetically complex compared to the dopaminergic system, it might be impossible to delineate a single functional domain for this neuromodulator [63]. Indeed the role of 5-HT might generalize to any sort of aversive signaling, including punishment (negative outcomes) but also effort (invigoration of action) and delay (opportunity cost).

Finally, the fourth hypothesis, which opposes structures such as the amygdala and anterior insula to the ventral striatum and prefrontal cortex, has been partially confirmed by investigations of patients with brain damage. Bilateral damage to the amygdala has been shown to impair implicit punishment learning, which was spared by bilateral damage to the hippocampus [64]. This classical observation has been more recently backed up by the finding that bilateral calcification in the amygdala abolishes loss aversion in economic decision-making [65]. To assess the implication of the insula in punishment avoidance, we administered to patients with brain tumors (Fig. 23.5A) the same task as that used with HD patients (Fig. 23.2B). Results (Fig. 23.5B) showed that insular lesions specifically impair punishment learning [66]. Computational analyses indicated that the deficit was best captured by decreasing the punishment parameter, which is consistent with fMRI studies reporting PPE encoding in the anterior insula [25,67,68]. Recent findings have shown the implication of the insula in effort learning, suggesting that this region might represent aversive signals across different domains [69].

Prolonging our investigation of HD using our probabilistic instrumental learning task (Fig. 23.3B), we found that at a symptomatic but still early stage of the disease, when neural degeneration affects both dorsal and ventral striatum, patients exhibited deficits in both punishment and reward learning [66]. The deficit in reward learning was best explained by reducing the reward parameter in the RL model. This aligns well with the position of the ventral striatum as a main output of the mesolimbic DA pathway, because DA drug effects were also captured by adjusting the reward parameter [25]. Results regarding the ventromedial prefrontal cortex (VMPFC) are not that clear-cut. Because activity in this region has been repeatedly shown to encode value positively, across both appetitive and aversive items [70,71], one would expect VMPFC damage to impair reward learning. Yet patients with VMPFC lesions were found to have more difficulty with learning from negative feedback [72] in a probabilistic learning task (Hiragana task, Fig. 23.2A). This could mean that contrary to the assumption of hypothesis 4 (Fig. 23.3A, bottom), the precision of coding in some cortical regions might actually be better for decreasing firing rates, unlike what was seen in neuromodulatory systems. Note

that in other studies from the same group, VMPFC patients also exhibited deficits in choosing between rewards [73].

CONCLUSIONS, LIMITATIONS, AND PERSPECTIVES

Whereas the implication of dopaminergic midbrain nuclei and ventral prefrontostriatal circuits in reward learning is quite well established, the delineation of an opponent system responsible for punishment learning is still a matter of debate. Our succinct review of the literature comparing reward and punishment learning in humans brings evidence to all four hypotheses regarding the neural implementation of a tentative punishment learning system. This could mean that various brain structures play a role in punishment learning: first those that were implicated in reward learning (DA, ventral striatum, VMPFC), second other neuromodulators such as 5-HT, and third other subcortical and cortical structures such as amygdala and anterior insula.

Yet this complicated picture might arise from some limitations in the approaches that were reviewed in this chapter. It is likely that the behavioral tasks do not purely target instrumental learning processes as implemented in RL models. Obviously a good fit of behavioral choices is no proof that the brain actually implements the computations operated in the models. Several learning systems might work in parallel to solve the choice problems, irrespective of outcome valence [74,75]. For instance, model-based and Pavlovian systems might interact with the model-free instrumental system that is formalized by the computational models commonly used to account for choice behavior. Another issue is that expected values are both subject- and context-dependent, which means that once state values are learned, some individuals might reframe their expectations (change their reference point), such that not winning can be perceived as a punishment, and not losing as a reward [76–78]. A related issue is that rewards and punishments are often confounded with positive and negative prediction errors. In fact, positive prediction errors can occur during punishment learning (when expected punishment is not received) and negative prediction errors during reward learning (when expected reward is not received). It is not clear at present whether the most relevant distinction for dividing brain systems is negative versus positive outcome (reward versus punishment) or negative versus positive prediction error. Finally, it is striking that most computationally characterized deficits were explained by changes in the outcome parameter (reward or punishment magnitude), or the choice temperature, but not the learning rate [25,43,51,66,79]. This leaves open the possibility that the effects of drugs and lesions reported here were not affecting learning processes per se, but biased other representations that had an impact on the learning plateau.

There are other limitations that are common to any drug or lesion study. Notably, pharmacological manipulations have tonic effects that only indirectly influence the phasic signals assumed to drive learning. Also, given the multiplicity of receptors and the complex interactions with genotypes, it is naïve to expect similar effects across subjects and across drugs that target the same neuromodulatory system. Nonetheless, characterizing deficits in computational terms might provide insight into pathological conditions and help predict the effects of treatment. For instance, PD patients who exhibit a strong increase in reward sensitivity following administration of a DA agonist might be at risk of developing an impulse control disorder [43] (see also Ref. [80] in the present volume). Computational analyses of learning abilities might also help us understand psychiatric diseases such as schizophrenia, which has been shown to impair processing of prediction errors and consequently lead to distorted representations of the environment [81].

References

[1] Barto AG, Sutton RS. Reinforcement learning: an introduction. Cambridge: MIT Press; 1998. http://dx.doi.org/10.1109/TNN.1998.712192.

[2] Rescorla RA, Wagner AR. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning II: current research and theory. New York: Applenton-Century-Crofts; 1972. p. 64–99.

[3] Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science 1997;275:1593–9. http://dx.doi.org/10.1126/science.275.5306.1593.

[4] Zaghloul KA, Blanco JA, Weidemann CT, McGill K, Jaggi JL, Baltuch GH, et al. Human substantia nigra neurons encode unexpected financial rewards. Science 2009;323:1496–9. http://dx.doi.org/10.1126/science.1167342.

[5] Fiorillo CD, Tobler PN, Schultz W. Discrete coding of reward probability and uncertainty by dopamine neurons. Science 2003;299:1898–902. http://dx.doi.org/10.1126/science.1077349.

[6] D'Ardenne K, McClure SM, Nystrom LE, Cohen JD. BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 2008;319:1264–7. http://dx.doi.org/10.1126/science.1150605.

[7] Haruno M, Kawato M. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. J Neurophysiol 2006;95:948–59. http://dx.doi.org/10.1152/jn.00382.2005.

[8] O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 2004;304:452–4. http://dx.doi.org/10.1126/science.1094285.

[9] Palminteri S, Boraud T, Lafargue G, Dubois B, Pessiglione M. Brain hemispheres selectively track the expected value of contralateral options. J Neurosci 2009;29:13465–72. http://dx.doi.org/10.1523/JNEUROSCI.1500-09.2009.
[10] Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 2005;47:129–41. http://dx.doi.org/10.1016/j.neuron.2005.05.020.
[11] Daw ND, Kakade S, Dayan P. Opponent interactions between serotonin and dopamine. Neural Networks 2002;15:603–16. http://dx.doi.org/10.1016/S0893-6080(02)00052-7.
[12] Guitart-Masip M, Duzel E, Dolan R, Dayan P. Action versus valence in decision making. Trends Cogn Sci 2014;18:194–202. http://dx.doi.org/10.1016/j.tics.2014.01.003. Elsevier Ltd.
[13] Seymour B, Maruyama M, De Martino B. When is a loss a loss? Excitatory and inhibitory processes in loss-related decision-making. Curr Opin Behav Sci 2015;5:122–7. http://dx.doi.org/10.1016/j.cobeha.2015.09.003.
[14] Pessiglione M, Delgado MR. The good, the bad and the brain: neural correlates of appetitive and aversive values underlying decision making. Curr Opin Behav Sci 2015;5:78–84. http://dx.doi.org/10.1016/j.cobeha.2015.08.006.
[15] Knutson B, Katovich K, Suri G. Inferring affect from fMRI data. Trends Cogn Sci 2014:1–7. http://dx.doi.org/10.1016/j.tics.2014.04.006. Elsevier Ltd.
[16] Schultz W. Electrophysiological correlates of reward processing in dopamine neurons. In: Léon T, Dreher J-C, editors. Decision neuroscience.
[17] Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 2011;14:154–62. http://dx.doi.org/10.1038/nn.2723. Nature Publishing Group.
[18] Matsumoto M, Hikosaka O. Lateral habenula as a source of negative reward signals in dopamine neurons. Nature 2007;447:1111–5. http://dx.doi.org/10.1038/nature05860.
[19] Lawson RP, Seymour B, Loh E, Lutti A, Dolan RJ, Dayan P, et al. The habenula encodes negative motivational value associated with primary punishment in humans. Proc Natl Acad Sci 2014. http://dx.doi.org/10.1073/pnas.1323586111.
[20] Brooks AM, Berns GS. Aversive stimuli and loss in the mesocorticolimbic dopamine system. Trends Cogn Sci 2013:1–6. http://dx.doi.org/10.1016/j.tics.2013.04.001. Elsevier Ltd.
[21] Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 2009;459:837–41. http://dx.doi.org/10.1038/nature08028. Macmillan Publishers Limited. All rights reserved.
[22] Seymour B, Daw N, Dayan P, Singer T, Dolan R. Differential encoding of losses and gains in the human striatum. J Neurosci 2007;27:4826–31. http://dx.doi.org/10.1523/JNEUROSCI.0400-07.2007.
[23] Delgado MR, Li J, Schiller D, Phelps EA. The role of the striatum in aversive learning and aversive prediction errors. Philos Trans R Soc Lond B Biol Sci 2008;363:3787–800. http://dx.doi.org/10.1098/rstb.2008.0161.
[24] Pauli WM, Larsen T, Collette S, Tyszka JM, Seymour B, O'Doherty JP. Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. J Neurosci 2015;35:14220–33. http://dx.doi.org/10.1523/JNEUROSCI.2277-15.2015.
[25] Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 2006;442:1042–5. http://dx.doi.org/10.1038/nature05051.
[26] Shenhav A, Buckner RL. Neural correlates of dueling affective reactions to win-win choices. Proc Natl Acad Sci USA 2014;111:10978–83. http://dx.doi.org/10.1073/pnas.1405725111.
[27] Soubrié P. Reconciling the role of central serotonin neurons in human and animal behavior. Behav Brain Sci 1986;9:319. http://dx.doi.org/10.1017/S0140525X00022871.
[28] Deakin JF, Graeff FG. 5-HT and mechanisms of defence. J Psychopharmacol 1991;5:305–15. http://dx.doi.org/10.1177/026988119100500414.
[29] Kapur S, Remington G. Serotonin-dopamine interaction and its relevance to schizophrenia. Am J Psychiatry 1996;153:466–76. http://dx.doi.org/10.1176/ajp.153.4.466.
[30] Lorrain DS, Riolo JV, Matuszewich L, Hull EM. Lateral hypothalamic serotonin inhibits nucleus accumbens dopamine: implications for sexual satiety. J Neurosci 1999;19:7648–52. Available: http://www.ncbi.nlm.nih.gov/pubmed/10460270.
[31] SB, Salzman CD. Appetitive and aversive systems in the amygdala. In: Tremblay L, Dreher J-C, editors. Decision neuroscience.
[32] Hayes DJ, Duncan NW, Xu J, Northoff G. A comparison of neural responses to appetitive and aversive stimuli in humans and other mammals. Neurosci Biobehav Rev 2014;45:350–68. http://dx.doi.org/10.1016/j.neubiorev.2014.06.018.
[33] Namburi P, Al-Hasani R, Calhoon GG, Bruchas MR, Tye KM. Architectural representation of valence in the limbic system. Neuropsychopharmacology 2015. http://dx.doi.org/10.1038/npp.2015.358.
[34] Bartra O, McGuire JT, Kable JW. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 2013;76:412–27. http://dx.doi.org/10.1016/j.neuroimage.2013.02.063. Elsevier Inc.
[35] Garrison J, Erdeniz B, Done J. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. Neurosci Biobehav Rev 2013:1–14. http://dx.doi.org/10.1016/j.neubiorev.2013.03.023. Elsevier Ltd.
[36] Yacubian J, Gläscher J, Schroeder K, Sommer T, Braus DF, Büchel C. Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J Neurosci 2006;26:9530–7. http://dx.doi.org/10.1523/JNEUROSCI.2915-06.2006.
[37] Büchel C, Dolan RJ. Classical fear conditioning in functional neuroimaging. Curr Opin Neurobiol 2000;10:219–23. http://dx.doi.org/10.1016/S0959-4388(00)00078-7.
[38] Frank MJ, Seeberger LC, Reilly RCO, O'Reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 2004;306:1940–3. http://dx.doi.org/10.1126/science.1102941.
[39] Samii A, Nutt JG, Ransom BR. Parkinson's disease. Lancet 2004;363:1783–93. http://dx.doi.org/10.1016/S0140-6736(04)16305-8.
[40] Frank MJ. Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw 2006;19:1120–36. http://dx.doi.org/10.1016/j.neunet.2006.03.006.
[41] Kéri S, Moustafa Aa, Myers CE, Benedek G, Gluck MA. α-Synuclein gene duplication impairs reward learning. Proc Natl Acad Sci USA 2010;107:15992–4. http://dx.doi.org/10.1073/pnas.1006068107. www.pnas.org/cgi/doi/10.1073/pnas.1006068107/-/DCSupplemental.
[42] Bódi N, Kéri S, Nagy H, Moustafa A, Myers CE, Daw N, et al. Reward-learning and the novelty-seeking personality: a between-and within-subjects study of the effects of dopamine agonists on young Parkinsons patients. Brain 2009;132:2385–95. http://dx.doi.org/10.1093/brain/awp094.
[43] Voon V, Pessiglione M, Brezing C, Gallea C, Fernandez HH, Dolan RJ, et al. Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. Neuron 2010;65:135–42. http://dx.doi.org/10.1016/j.neuron.2009.12.027. Elsevier Inc.
[44] Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, Ullsperger M. Genetically determined differences in learning from errors. Science 2007;318:1642–5. http://dx.doi.org/10.1126/science.1145044. American Association for the Advancement of Science.

[45] Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci USA 2007;104:16311–6. http://dx.doi.org/10.1073/pnas.0706111104.
[46] Pessiglione M, Petrovic P, Daunizeau J, Palminteri S, Dolan RJ, Frith CD. Subliminal instrumental conditioning demonstrated in the human brain. Neuron 2008;59:561–7. http://dx.doi.org/10.1016/j.neuron.2008.07.005.
[47] Palminteri S, Lebreton M, Worbe Y, Grabli D, Hartmann A, Pessiglione M. Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes. Proc Natl Acad Sci USA 2009;106:19179–84. http://dx.doi.org/10.1073/pnas.0904035106.
[48] Leckman JF. Tourette's syndrome. Lancet 2002;360:1577–86. http://dx.doi.org/10.1016/S0140-6736(02)11526-1.
[49] Rutledge RB, Lazzaro SC, Lau B, Myers CE, Gluck MA, Glimcher PW. Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. J Neurosci 2009;29:15104–14. http://dx.doi.org/10.1523/JNEUROSCI.3524-09.2009.
[50] Jocham G, Klein Ta, Ullsperger M. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. J Neurosci 2011;31:1606–13. http://dx.doi.org/10.1523/JNEUROSCI.3904-10.2011.
[51] Eisenegger C, Naef M, Linssen A, Clark L, Gandamaneni PK, Mu U. Role of dopamine D2 receptors in human reinforcement learning. 2014. p. 1–10. http://dx.doi.org/10.1038/npp.2014.84.
[52] Walker FO. Huntington's disease. Semin Neurol 2007;27:143–50. http://dx.doi.org/10.1055/s-2007-971176.
[53] Douaud G, Gaura V, Ribeiro M-J, Lethimonnier F, Maroy R, Verny C, et al. Distribution of grey matter atrophy in Huntington's disease patients: a combined ROI-based and voxel-based morphometric study. Neuroimage 2006;32:1562–75. http://dx.doi.org/10.1016/j.neuroimage.2006.05.057.
[54] den Ouden HEM, Daw ND, Fernandez G, Elshout Ja, Rijpkema M, Hoogman M, et al. Dissociable effects of dopamine and serotonin on reversal learning. Neuron 2013;80:1090–100. http://dx.doi.org/10.1016/j.neuron.2013.08.030.
[55] Guitart-Masip M, Economides M, Huys QJM, Frank MJ, Chowdhury R, Duzel E, et al. Differential, but not opponent, effects of l-DOPA and citalopram on action learning with reward and punishment. Psychopharmacol (Berl) 2014;231:955–66. http://dx.doi.org/10.1007/s00213-013-3313-4.
[56] Cools R, Robinson OJ, Sahakian B. Acute tryptophan depletion in healthy volunteers enhances punishment prediction but does not affect reward prediction. Neuropsychopharmacology 2008;33:2291–9. http://dx.doi.org/10.1038/sj.npp.1301598.
[57] Seymour B, Daw NDN, Roiser JPJ, Dayan P, Dolan R. Serotonin selectively modulates reward value in human decision-making. J Neurosci 2012;32:5833–42. http://dx.doi.org/10.1523/JNEUROSCI.0053-12.2012.
[58] Palminteri S, Clair A-HH, Mallet L, Pessiglione M. Similar improvement of reward and punishment learning by serotonin reuptake inhibitors in obsessive-compulsive disorder. Biol Psychiatry 2012;72:244–50. http://dx.doi.org/10.1016/j.biopsych.2011.12.028. Elsevier Inc.
[59] Crockett MJ, Clark L, Robbins TW. Reconciling the role of serotonin in behavioral inhibition and aversion: acute tryptophan depletion abolishes punishment-induced inhibition in humans. J Neurosci 2009;29:11993–9. http://dx.doi.org/10.1523/JNEUROSCI.2513-09.2009.
[60] Boureau Y-L, Dayan P. Opponency revisited: competition and cooperation between dopamine and serotonin. Neuropsychopharmacology 2011;36:74–97. http://dx.doi.org/10.1038/npp.2010.151. Nature Publishing Group.
[61] Tanaka SC, Schweighofer N, Asahi S, Shishida K, Okamoto Y, Yamawaki S, et al. Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum. PLoS One 2007;2:e1333. http://dx.doi.org/10.1371/journal.pone.0001333.
[62] Schweighofer N, Bertin M, Shishida K, Okamoto Y, Tanaka SC, Yamawaki S, et al. Low-serotonin levels increase delayed reward discounting in humans. J Neurosci 2008;28:4528–32. http://dx.doi.org/10.1523/JNEUROSCI.4982-07.2008.
[63] Spies M, Knudsen GM, Lanzenberger R, Kasper S. The serotonin transporter in psychiatric disorders: insights from PET imaging. The Lancet Psychiatry 2015;2:743–55. http://dx.doi.org/10.1016/S2215-0366(15)00232-1.
[64] Bechara A, Tranel D, Damasio H, Adolphs R, Rockland C, Damasio A. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. Science 1995;269:1115–8. http://dx.doi.org/10.1126/science.7652558.
[65] De Martino B, Camerer CF, Adolphs R. Amygdala damage eliminates monetary loss aversion. Proc Natl Acad Sci USA 2010;107:3788–92. http://dx.doi.org/10.1073/pnas.0910230107.
[66] Palminteri S, Justo D, Jauffret C, Pavlicek B, Dauta A, Delmaire C, et al. Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. Neuron 2012;76:998–1009. http://dx.doi.org/10.1016/j.neuron.2012.10.017.
[67] Kim H, Shimojo S, O'Doherty JP. Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. PLoS Biol 2006;4:e233. http://dx.doi.org/10.1371/journal.pbio.0040233.
[68] Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, et al. Temporal difference models describe higher-order learning in humans. Nature 2004;429:664–7. http://dx.doi.org/10.1038/nature02636.1.
[69] Skvortsova V, Palminteri S, Pessiglione M. Learning to minimize efforts versus maximizing rewards: computational principles and neural correlates. J Neurosci 2014;34:15621–30. http://dx.doi.org/10.1523/JNEUROSCI.1350-14.2014.
[70] Tom SM, Fox CR, Trepel C, Poldrack RA. The neural basis of loss aversion in decision-making under risk. Science 2007;315:515–8. http://dx.doi.org/10.1126/science.1134239.
[71] Plassmann H, O'Doherty JP, Rangel A. Appetitive and aversive goal values are encoded in the medial orbitofrontal cortex at the time of decision making. J Neurosci 2010;30:10799–808. http://dx.doi.org/10.1523/JNEUROSCI.0788-10.2010.
[72] Wheeler EZ, Fellows LK. The human ventromedial frontal lobe is critical for learning from negative feedback. Brain 2008;131:1323–31. http://dx.doi.org/10.1093/brain/awn041.
[73] Camille N, Griffiths Ca, Vo K, Fellows LK, Kable JW. Ventromedial frontal lobe damage disrupts value maximization in humans. J Neurosci 2011;31:7527–32. http://dx.doi.org/10.1523/JNEUROSCI.6527-10.2011.
[74] Daw ND. Advanced reinforcement learning. In: Glimcher PW, Fehr E, editors. Neuroeconomics decis mak brain. 2nd ed. London, UK: Neuroecono. Academic Press; 2013. p. 299–320. http://dx.doi.org/10.1016/B978-0-12-416008-8.00016-4.
[75] Dayan P. Twenty-five lessons from computational neuromodulation. Neuron 2012;76:240–56. http://dx.doi.org/10.1016/j.neuron.2012.09.027. Elsevier Inc.
[76] Vlaev I, Chater N, Stewart N, Brown GD. Does the brain calculate value? Trends Cogn Sci 2011;15:546–54. http://dx.doi.org/10.1016/j.tics.2011.09.008. Elsevier Ltd.
[77] Seymour B, McClure SM. Anchors, scales and the relative coding of value in the brain. Curr Opin Neurobiol 2008;18:173–8. http://dx.doi.org/10.1016/j.conb.2008.07.010.
[78] Rangel A, Clithero Ja. Value normalization in decision making: theory and evidence. Curr Opin Neurobiol 2012;22:970–81. http://dx.doi.org/10.1016/j.conb.2012.07.011. Elsevier Ltd.

[79] Shiner T, Seymour B, Wunderlich K, Hill C, Bhatia KP, Dayan P, et al. Dopamine and performance in a reinforcement learning task: evidence from Parkinson's disease. Brain 2012;135: 1871–83. http://dx.doi.org/10.1093/brain/aws083.

[80] Voon V. Decision making and impulse control disorders in Parkinson's disease. In: Tremblay L, Dreher J-C, editors. Decision neuroscience.

[81] Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci 2009;10:48–58. http://dx.doi.org/10.1038/nrn2536.

[82] Hull CL. Principles of behavior: an introduction to behavior theory [internet]. 1943. Available: https://books.google.fr/books/about/Principles_of_Behavior.html?id=6WB9AAAAMAAJ&pgis=1.

[83] Watkins CJCH, Dayan P. Q-learning. Mach Learn 1992;8:279–92. http://dx.doi.org/10.1007/BF00992698.

[84] Luce RD. Individual choice behavior: a theoretical analysis [internet]. Courier Corporation; 1959. Available: https://books.google.com/books?hl=en&lr=&id=D74qAwAAQBAJ&pgis=1.

[85] O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. Ann NY Acad Sci 2007;1104:35–53. http://dx.doi.org/10.1196/annals.1390.022.

[86] Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci 2012;16:72–80. http://dx.doi.org/10.1016/j.tics.2011.11.018.

[87] Palminteri S, Khamassi M, Joffily M, Coricelli G. Contextual modulation of value signals in reward and punishment learning. Nat Commun 2015;6:8096. http://dx.doi.org/10.1038/ncomms9096. Nature Publishing Group.

[88] Guitart-Masip M, Fuentemilla L, Bach DR, Huys QJM, Dayan P, Dolan RJ, et al. Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. J Neurosci 2011;31:7867–75. http://dx.doi.org/10.1523/JNEUROSCI.6376-10.2011.

[89] Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, et al. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput Biol 2011;7:e1002028. http://dx.doi.org/10.1371/journal.pcbi.1002028.

[90] Yarkoni T, Poldrack Ra, Nichols TE, Van Essen DC, Wager TD. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 2011;8:665–70. http://dx.doi.org/10.1038/nmeth.1635.

[91] Singer HS. Tourette's syndrome: from behaviour to biology. Lancet Neurol 2005;4:149–59. http://dx.doi.org/10.1016/S1474-4422(05)01012-4.

[92] Palminteri S, Pessiglione M. Reinforcement learning and tourette syndrome [internet]. In: International review of neurobiology. 1st ed. Elsevier Inc.; 2013. http://dx.doi.org/10.1016/B978-0-12-411546-0.00005-6.