PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Modeling the effects of motivation on choice and learning in the basal ganglia

Maaike M. H. van Swieten, Rafal Bogacz*

MRC Brain Network Dynamics Unit, University of Oxford, Oxford, United Kingdom

* rafal.bogacz@ndcn.ox.ac.uk

Abstract

Decision making relies on adequately evaluating the consequences of actions on the basis of past experience and the current physiological state. A key role in this process is played by the basal ganglia, where neural activity and plasticity are modulated by dopaminergic input from the midbrain. Internal physiological factors, such as hunger, scale signals encoded by dopaminergic neurons and thus they alter the motivation for taking actions and learning. However, to our knowledge, no formal mathematical formulation exists for how a physiological state affects learning and action selection in the basal ganglia. We developed a framework for modelling the effect of motivation on choice and learning. The framework defines the motivation to obtain a particular resource as the difference between the desired and the current level of this resource, and proposes how the utility of reinforcements depends on the motivation. To account for dopaminergic activity previously recorded in different physiological states, the paper argues that the prediction error encoded in the dopaminergic activity needs to be redefined as the difference between utility and expected utility, which depends on both the objective reinforcement and the motivation. We also demonstrate a possible mechanism by which the evaluation and learning of utility of actions can be implemented in the basal ganglia network. The presented theory brings together models of learning in the basal ganglia with the incentive salience theory in a single simple framework, and it provides a mechanistic insight into how decision processes and learning in the basal ganglia are modulated by the motivation. Moreover, this theory is also consistent with data on neural underpinnings of overeating and obesity, and makes further experimental predictions.

Author summary

Behaviour is made of decisions that are based on the evaluation of costs and benefits of potential actions in a given situation. Actions are usually generated in response to reinforcement cues which are potent triggers of desires that can range from normal appetites to compulsive addictions. However, learned cues are not constant in their motivating power. Food cues are more potent when you are hungry than when you have just finished a meal. These changes in cue-triggered desire produced by a change in biological state present a challenge to many current computational models of motivation and learning. Here, we demonstrate concrete examples of how motivation can instantly modulate

OPEN ACCESS

Citation: van Swieten MMH, Bogacz R (2020) Modeling the effects of motivation on choice and learning in the basal ganglia. PLoS Comput Biol 16(5): e1007465. https://doi.org/10.1371/journal.pcbi.1007465

Editor: Jonathan Rubin, University of Pittsburgh, UNITED STATES

Received: October 3, 2019

Accepted: April 3, 2020

Published: May 26, 2020

Data Availability Statement: All relevant data are within the manuscript (this is a modelling study).

Funding: MMHvS was supported by Medical Research Council (mrc.ukri.org) studentship MC_ST_U16043. RB was supported by Medical Research Council (mrc.ukri.org) grant MC_UU_12024/5 and Biotechnology and Biological Sciences Research Council (bbsrc.ukri.org) grant BB/S006338/1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

reinforcement values and actions; we propose an overarching framework of learning and action selection based on maintaining the physiological balance to better capture the dynamic interaction between learning and physiology that controls the incentive salience mechanism of motivation for reinforcements. These models provide a unified account of state-dependent learning of the incentive value of actions and selecting actions according to the learned positive and negative consequences of those actions and with respect to the physiological state. We propose a biological implementation of how these processes are controlled by an area in the brain called the basal ganglia, which is associated with error-driven learning.

Introduction

Successful interactions with the environment rely on using previous experience to predict the value of outcomes or consequences of available actions. Human and animal studies have strongly implicated the neurotransmitter dopamine in these learning processes [1–8], in addition to its roles in shaping behaviour, including motivation [9], vigour [10] and behavioural activation [11, 12].

Dopamine seems to have two distinct effects on the networks it modulates. First, it facilitates learning by triggering synaptic plasticity [13]. Such dopaminergic teaching signal is thought to encode a reward prediction error (RPE), which is defined as a difference between a reinforcement and the expected reinforcement [1]. The overall value of a reinforcement that is available at a given moment depends on the potential positive and negative consequences associated with obtaining it. These consequences can be influenced by internal and external factors, such as the physiology of the subject and the reinforcement's availability, respectively. Information about the external factors is indeed encoded in the dopaminergic responses which are shown to scale with the magnitude and the probability of a received reinforcement [14, 15], but also with the delay and effort related costs associated with a reinforcement [16, 17]. Second, the level of dopamine controls the activation of the basal ganglia network by modulating the excitability of neurons [18–20]. Although dopamine is a critical modulator of both learning and activation, it is unclear how it is able to do both given that these processes are conceptually, computationally and behaviourally distinct. For a long time, our understanding was that tonic (sustained) levels of dopamine encode an activation signal and phasic (transient) responses convey a teaching signal (i.e. prediction error) [10]. However, recent studies have shown that this distinction is not as clear as we thought [21, 22] and that other mechanisms may exist, which allow striatal neurons to correctly decode the two signals from dopaminergic activity [23]. In this paper, we do not investigate mechanisms by which these different signals can be accessed, but we assume that striatal neurons can read out both activation and teaching signals encoded by dopaminergic neurons.

In addition to the external factors explained above, internal physiological factors, such as hunger, can also alter the reinforcement value of an action and drive decision making based on the usefulness of that action and the outcome at that given time. For example, searching for food when hungry is more valuable than when sated and actions have to be evaluated accordingly. In other words, the current physiological state affects the motivation to obtain a particular resource. The physiological state has indeed been observed to modulate dopamine levels and dopamine responses encoding reward prediction error [24–26], thus it is likely that the physiological state influences both the activation and teaching signals carried by dopamine. Strikingly, the physiological state can sometimes even reverse the value of a reinforcement (e.g.

salt) from being rewarding in a depleted state to aversive in a sated state [27]. Moreover, the physiological state during learning may affect subsequent choices, for example, animals may still have a preference for actions that were previously associated with hunger even when they are sated [28]. Recently, it has been proposed how a physiological state can be introduced into reinforcement learning theory to refine the definition of a reinforcement [29]. However, despite the importance of the physiological state for describing behaviour and dopaminergic activity, we are not aware of theoretical work that integrates the physiological state into a theory of dopaminergic responses.

Another important line of work describing subjective preferences is the utility theory. It is based on the assumption that people can consistently rank their choices depending upon their preferences. The utility theory has been used extensively in economics [30], and it has been shown that dopaminergic responses depend on the subjective utility of the obtained reward magnitude, rather than its objective magnitude [31]. As described above, there is a need to extend the general utility function with a motivational component that describes the bias in the evaluation of positive and negative consequences of decisions as a result of changes in the physiological state of a subject. Evidence for this bias comes from devaluation studies in which reinforcements are specifically devalued by pre-feeding or taste aversion. The concept of state-dependent valuation has been studied in various contexts [25, 32, 33] and in different species, including starlings [34, 35], locusts [36] and fish [28]. These studies suggest that the utility of outcomes depends on both the (learned) reinforcement value and the physiological state. One of the earliest attempts to capture this relationship between incentive value and internal motivational state is the incentive salience theory [12].

In this paper we aim to provide an explanation for the above effects of physiological state on behaviour and dopaminergic activity with a simple framework that combines incentive learning theory [37, 38] with models of learning in the basal ganglia. By integrating key concepts from these theories we define a utility function for actions that can be modulated by internal and external factors. In our framework, the utility is defined as the change in the desirability of physiological state resulting from taking an action and obtaining a reinforcement. Following previous theoretic work [29], the motivation for a particular resource is defined as the difference between the desired and the current level of this resource.

In the proposed framework, motivation affects both teaching and activation signals encoded by dopaminergic neurons. Relying on experimental data, we argue that the dopaminergic teaching signal encodes the difference between utility and expected utility, which depends on motivation. Moreover, we propose how motivation can influence the dopaminergic activation signal to appropriately drive action selection behaviour. We also highlight that the resulting consequences of an action can be positive or negative depending on how far the current and new physiological state are from the desired state. Building on existing theories we illustrate how the neurons in the striatum could learn these consequences through plasticity rules. Finally, we use the resulting models to explain experimental data. Together, this paper discusses a modelling framework that describes how the internal physiological state affects learning and action selection in the basal ganglia and provides novel interpretations of existing experimental data. To provide a rationale for our framework the remainder of the introduction reviews the data on effects of physiological state on dopaminergic teaching signal.

Effects of motivation on dopaminergic responses

We first review a classical reinforcement learning theory and then discuss data that challenges it. As postulated by reinforcement learning theories, expectations of outcomes are updated on the basis of experiences. This updating process may be guided by prediction errors, which are

computed by subtracting the cached reinforcement expectation (V_t) from the received reinforcement (r). In classical conditioning and after extensive training, the dopamine response to the conditioned stimulus (CS) is observed to reflect the expected future reinforcement, whereas the response to the unconditioned stimulus (US) represents the difference between the obtained reinforcement and the expectation [1]. To account for these responses, the reward prediction error in a temporal difference model (δ_{TD}) is classically defined as [1]:

$$\delta_{TD} = r_t + V_{t+1} - V_t. \quad (1)$$

The above equation defines the prediction error as the difference between total reinforcement (including both reinforcement actually received r_t and reinforcement expected in the future V_{t+1}) and the expected reinforcement (V_t).

We now review how the above equation captures the dopamine responses at the time of the CS and the US, which change over the course of learning. At the start of learning, the animal has not formed any expectation yet, which means that at the time of the CS, V_t is 0. Given that no reinforcement is provided at the time of CS presentation, r_t is also 0. Thus, the prediction error at the time of the CS is equal to the expected value of the reinforcement ($\delta_{TD} = V_{t+1}$). The response to the CS is zero in naive animals. By contrast, the response to the CS in fully trained animals reflects the expected upcoming reinforcement, as extensive training allowed animals to update their expectations to predict upcoming reinforcements better. At the time of the US, no future reinforcements are expected so V_{t+1} is 0, thus the reward prediction error at the time of US is equal to $\delta_{TD} = r_t - V_t$. Unpredicted rewards (i.e. positive reinforcements) evoke positive prediction errors, while predicted reinforcements do not. This definition of the prediction error captures observed patterns of dopaminergic responses; where naive animals, which are unable to predict reinforcements, show large positive responses at the time of the US, and fully trained animals, which can predict reinforcements perfectly, show no response at all.

Recently, the above definition of reward prediction error has been experimentally challenged by Cone et al. [26]. They show that the internal state of an animal modulates the teaching signals encoded by dopamine neurons in the midbrain after conditioning (Fig 1). In this study, animals were trained and tested in either a sodium depleted or sodium balanced state. The dopaminergic responses predicted by the classical reinforcement learning theory shown

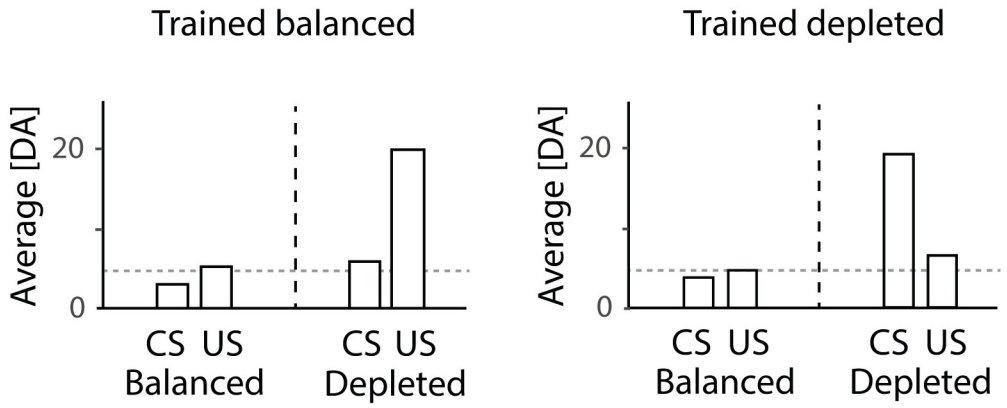

Fig 1. State-dependent modulation of dopaminergic responses. Experimental data by Cone et al. [26] shows dynamic changes in dopaminergic responses based on the state of the animal. The two graphs within the figure correspond to dopaminergic responses in animals trained in a balanced and depleted state, respectively, re-plotted from figures 2 and 4 in the paper by Cone et al. [26]. Within each graph, the left and right halves show the responses of animals tested in balanced and depleted states, respectively. The horizontal dashed lines indicate baseline levels.

https://doi.org/10.1371/journal.pcbi.1007465.g001

by Schultz et al. [1] were only observed in animals that were both trained and tested in the depleted state. In all other conditions the dopaminergic responses followed different patterns. When animals were trained in the balanced state but tested in the depleted state, increased dopaminergic responses to the US (i.e. salt infusion) rather than the CS were observed, which is similar to dopaminergic responses observed in untrained animals in the study of Schultz et al. [1], suggesting that learning did not occur in the balanced state. When animals were trained and tested in the balanced state, there was no dopaminergic response to either CS or US. Interestingly, the same pattern was observed in animals trained in the depleted state and tested in the balanced state, suggesting that the learned values were modulated by the state at testing. In the Results section below we will demonstrate how these patterns of activity can be captured by appropriately modifying a definition of prediction error.

Results

In this section, we present our framework and its possible implementation in the basal ganglia circuit, and illustrate how it can account for the effects of motivation on neural activity and behaviour.

Normative theory of state-dependent utility

To gain intuition for how actions can be taken based on a state-dependent utility, we first develop a normative theory. The utility and consequences of actions are dependent on the usefulness of the reinforcement (r) with respect to the current state. Taking actions and obtaining reinforcements allows us to maintain our physiological balance by minimising the distance between the current state S and the desired state S^*. We assume that the desirability function of a physiological state has a concave, quadratic shape (Fig 2A), because it is more important to act when you are in a very low physiological state, compared to when you are in a near optimal state. Thus, we define the desirability of a state in the following way (a constant of 1/2 is added for mathematical convenience, as it will cancel in subsequent derivations):

$$Y(S) = -\frac{1}{2}(S - S^*)^2. \tag{2}$$

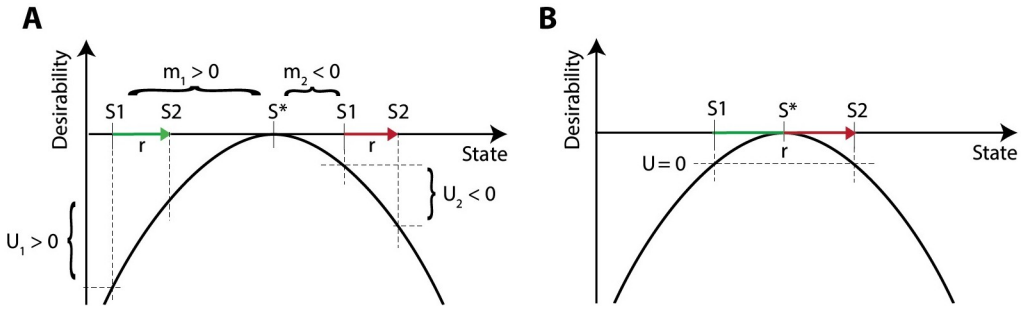

Fig 2. The utility of an action depends on the reinforcement size and physiological state. A) The same reinforcement can yield a positive or negative utility depending on whether the difference between the current and new physiological state is positive or negative. B) A large reinforcement may have an utility of zero even if the animal was initially in a depleted state. U = utility, m = motivation. S^* = the desired state and S1 = state before action, S2 = state after action. Arrow length indicates the size of the reinforcement (r). Changes in state resulting in an increase and decrease in desirability are indicated with green and red arrows, respectively.

https://doi.org/10.1371/journal.pcbi.1007465.g002

We define the motivation at a given physiological state as the difference between the desired and the current physiological state:

$$m = S^* - S. \tag{3}$$

For the quadratic desirability function defined in Eq (2), the motivation at any state S is also equal to the slope of the desirability function at this state $Y'(S) = m$. We define the utility U of an action that changes the state of the animal from $S1$ to $S2$ as the difference between the desirability at state 2, $Y(S2)$ and state 1, $Y(S1)$. A simple approximation for the utility can be obtained through a first order Taylor expansion of Eq (2):

$$U = Y(S2) - Y(S1) \tag{4}$$

$$\approx Y'(S1)(S2 - S1) \tag{5}$$

$$= mr. \tag{6}$$

This approximation of the utility clearly shows that the utility of an action, defined as the change in physiological state, depends on both the motivation m (Eq (3)) and the reinforcement r, where $r = S2 - S1$.

Moreover, according to the above definition, the same reinforcement could yield a positive or negative utility of an action, depending on whether the difference between the current physiological state and the desired state (i.e. the motivation m) is positive or negative (Fig 2A). This parallels an observation that nutrients such as salt may be appetitive or aversive depending on the level of an animal's reserves [27]. Although not discussed in this paper, please note that this definition of the utility also can be extended to the utility of an external state, such as a particular location in space. The utility of such an external state can be defined as a utility of the best available action in this state.

In order to select actions on the basis of their utility, animals need to maintain an estimate of the utility $\hat{U}$ of an action. There are several ways such an estimate can be learned. Here we discuss a particular learning algorithm, which results in prediction errors that resemble those observed by Cone et al. [26]. This learning algorithm assumes that animals minimise the absolute error in the prediction of the utility of the chosen action. We can therefore define this prediction error as:

$$\delta = U - \hat{U}. \tag{7}$$

The above expression for the prediction error (Eq (7)) provides a general definition of the prediction error as the difference between the observed and expected utility. In this paper we claim that this expression better describes the dopaminergic teaching signal observed in experimental data, which we will demonstrate in more detail in the next section.

Assuming that the animal's estimate of expected reinforcement is encoded in a parameter V, the animal's estimate of the utility is $\hat{U} = mV$. Combining Eq (6) with Eq (7), we obtain the following expression for the reward prediction error:

$$\delta = mr - mV. \tag{8}$$

A common technique to obtain better predictions is to minimise the absolute error, where the error is defined as the difference between the actual and the predicted utility as described in the above equation. For this optimisation method we use a gradient ascent and define an

objective function F that we will maximise:

$$F = -\frac{1}{2}\delta^2. \tag{9}$$

To increase this objective function, the estimate of the expected reinforcement, V, is changed in the direction of the gradient (i.e. the direction that most increases F), which results in an update proportional to the prediction error:

$$\Delta V \sim \frac{\partial F}{\partial V} = m\delta. \tag{10}$$

Simulating state-dependent dopaminergic responses

This section serves to illustrate that the pattern of dopaminergic activity seen in the study by Cone et al. [26] is not consistent with the classical theory and can be better explained with a state-dependent utility as described above. We first simulated the classical model in which reward prediction error is described in Eq (1). In the simulation, the CS was presented at time step 1, while the US was presented at time step 2. The model learned a single parameter V that estimates the expected reinforcement on the time step following the CS. Thus, on each trial the prediction error to the CS was equal to $\delta_{TD} = V$, while the prediction error to the US was equal to $\delta_{TD} = r - V$. The value estimate was updated proportionally to the prediction error, i.e. $\Delta V = \alpha (r - V)$, where α is a learning rate parameter. On every trial, the model received a reinforcement $r = 0.5$. Once training was completed, the expected values were fixed to the values they converged to during training, and testing occurred without allowing the model to update its beliefs.

The classical reward prediction error Eq (1) does not depend on the physiological state, therefore each experimental condition was simulated in exactly the same way. The dopaminergic teaching signals, predicted by the classical theory, are thus identical in all conditions (Fig 3A). As described above, the response to the CS is equal to the estimate of the reinforcement and the response to the US is equal to the difference between the received and the predicted reinforcement, which is close to zero when reinforcements are fully predicted.

Please note that the classical reward prediction error employed in the simulations shown in Fig 3A is identical to that in Eq (8) with $m = 1$. To achieve a state-dependent modulation we use the state-dependent prediction error (Eq (8)) to simulate the data by Cone et al. following the same protocol as in the classical case. For these simulations, the parameter describing motivation was set to $m = 0.2$ for a state close to balanced and $m = 2$ for a depleted state. During training in the depleted state, the reward prediction error to the US (Eq (8)) is high on the initial trials to facilitate learning of a reward estimate. By contrast, during training in the near-balanced state, the prediction errors are close to 0, and consequently the estimate of reinforcement is only minimally adjusted. In other words, the learned CS value is higher for the animals trained in the depleted state than the CS value of animals trained in the balanced state. During testing, these learned CS values were no longer modified. The dopaminergic teaching signal at the time of the US was computed from Eq (8) using the value m of the testing state. The response to the CS was taken as mV, i.e. motivation m of testing times the CS value learned during training.

In simulated animals that are trained in the near-balanced state little learning is triggered and the response to the CS is close to zero (Fig 3B). However, when these simulated animals are then tested in the depleted state, the scaled utility is greater than zero and consequently evokes a positive reward prediction error. In contrast, simulated animals trained in the depleted state learn the estimate of the expected value of the reinforcement. There is an

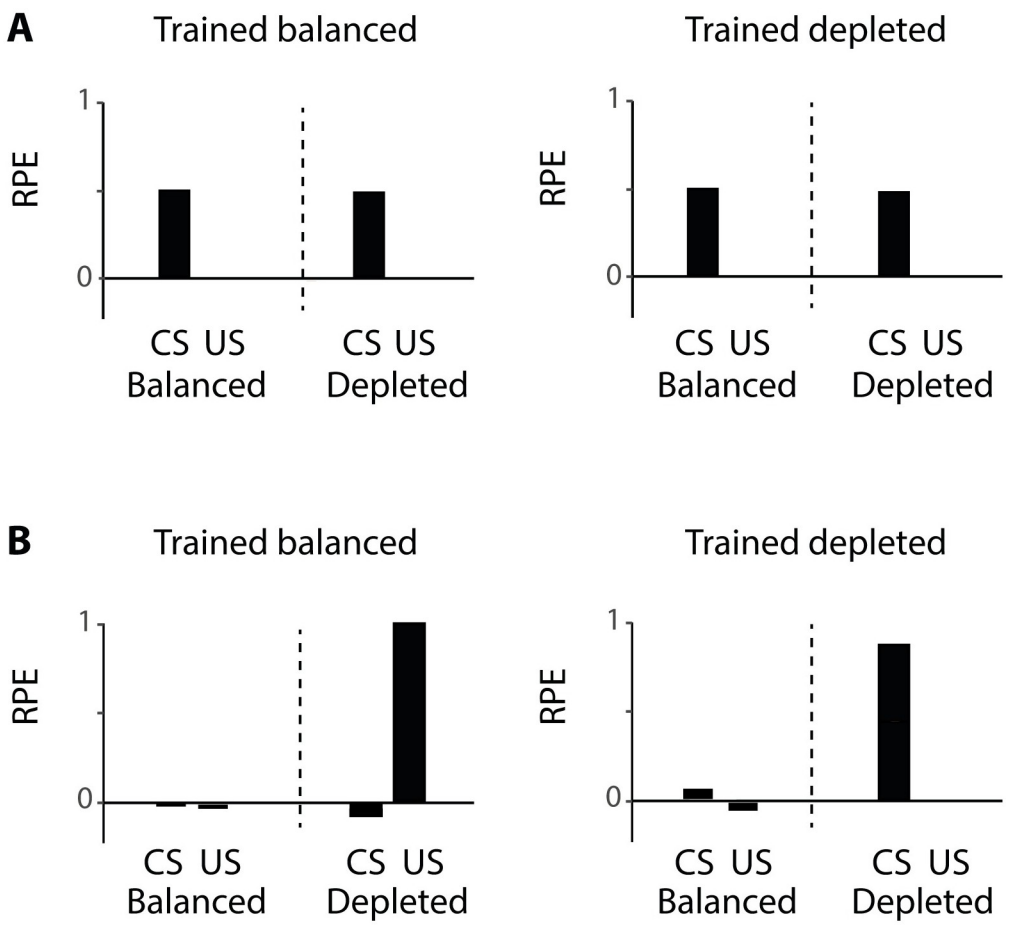

Fig 3. Simulation of data by Cone et al. [26] with different reward prediction errors. A) Simulation with classical reward prediction error using Eq (1). B) Simulation with state-dependent prediction error described in Eq (8). Within each graph, the left and right halves show the reward prediction error (RPE) of simulated animals tested in balanced and depleted states, respectively. CS = conditioned stimulus, US = unconditioned stimulus. Each simulation consisted of 50 training trials, 1 test trial and was repeated 5 times, similar to the number of animals in each group in the study by Cone et al. [26]. Error bars are equal to zero as there is no noise added to the simulation and all simulations converged to the same value.

https://doi.org/10.1371/journal.pcbi.1007465.g003

increase in the dopaminergic teaching signal in these simulated animals at the time of the CS since the expected value is transferred to the CS. When these simulated animals are tested in the near-balanced state, with a motivation close to zero, a very small reward prediction error is evoked, because both the reinforcement and expected value are scaled by a number close to zero.

In line with the theory in the previous section in which we formally defined both the utility and motivation, the above simulations shows that in order to account for the experimental data by Cone et al., (2016), the prediction error needs to be redefined as a difference between the utility of a reinforcement and the expected utility of that reinforcement, which depends on both the objective reinforcement magnitude and the motivation.

Accounting for positive and negative consequences of actions

Let us now reconsider the dependence of utility on motivation and consider how this could be expressed more accurately. Since Eq (4) comes from Taylor expansion, it only provides a close

approximation if r is small. This approximation may fail when the reinforcement is greater than the distance to the optimum. In the example in Fig 2B, if we use a linear approximation with a positive motivation, the utility is approximated as greater than zero, even though the actual utility is not as this action will exceed the desired state. Eq (4) also suggests that any action with $r > 0$ will have positive utility if $m > 0$, regardless of possible negative consequences (i.e. reaching a new state further away from the desired state). Moreover, if the distance of the current state to the desired state is equal to the distance of the new state to the desired state, the utility of an action would be zero (Fig 2B). Using Eq (4) it is impossible to capture these effects and account for both positive and negative consequences of this action.

One classical example in which the utility of an action switches sign depending on the proximity to the desired physiological state is salt appetite. When animals are depleted of sodium, salt consumption is rewarding. However, when animals are physiologically balanced, salt consumption is extremely aversive [27]. To avoid using multiple equations to explain the switch from positive to negative utilities and vice versa [38], we need to formulate an equation that can account for negative consequences of actions when the $m \approx 0$ or $m < 0$ and is able to account for positive consequences when the motivation changes, i.e. $m > 0$.

Therefore we use a second order Taylor expansion which includes the first and second order derivatives of the desirability function (Eq 2) and gives an exact expression for the utility:

$$U = Y(S2) - Y(S1) \tag{11}$$

$$= Y'(S1)(S2 - S1) + Y''(S1)(S2 - S1)^2/2 \tag{12}$$

$$= mr - r^2/2. \tag{13}$$

In the above equation mr could be seen as the positive and $r^2/2$ as the negative consequences of the action, respectively. The first term plays a greater role when deprived and promotes taking actions, whereas the second term plays a greater role when balanced and discourages taking actions.

During action selection it is imperative to choose actions to maximise future utility. For competing actions, the utility of all available actions needs to be computed. The action with the highest utility is most beneficial to select, but this action should only be chosen when its utility is positive. If the utility of all actions is negative, no actions should be taken. From a fitness point of view, not making an action is more advantageous than incurring a high cost.

In the next section we will elaborate on how Eq (13) can be evaluated in the basal ganglia and provide an example of a biologically plausible implementation. For simplicity we will only consider a single physiological dimension (e.g. nutrient reserve), but we recognise that the theory needs to be extended in the future to multiple dimensions (e.g. water reserve, fatigue) that animals needs to optimise. Furthermore, taking an action aimed to restore one dimension (e.g. nutrient reserve) may also include negative consequences that are independent of the considered dimension (e.g. fatigue). We will elaborate on these issues in the Discussion.

Neural implementation

In the previous sections we discussed how the utility of actions or stimuli change in a state-dependent manner. In this section we will focus on the neural implementation of these concepts. More specifically, we will address how the utility of previously chosen actions can be computed in the basal ganglia and how this circuit could learn the utility of actions.

Evaluation of utility in the basal ganglia circuit. The basal ganglia is a group of subcortical nuclei that play a key role in action selection and reinforcement learning. It is organised into two main pathways shown schematically in Fig 4. The Go or direct pathway is associated

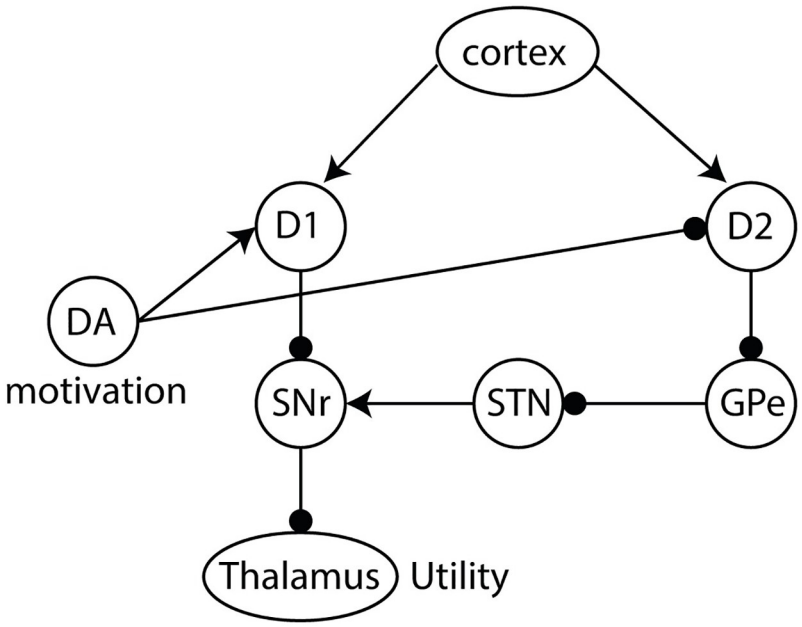

Fig 4. Schematic of utility computation in the basal ganglia network. Dopaminergic activation signal encodes the motivation. The thalamic activity represents the utility of actions. Arrows and lines with circles denote excitatory and inhibitory connections, respectively. DA = dopamine, D1 = dopamine receptor 1 medium spiny neurons, D2 = dopamine receptor 2 medium spiny neurons, SNr = substantia nigra pars reticulata, STN = subthalamic nucleus, GPe = globus pallidus external segment.

https://doi.org/10.1371/journal.pcbi.1007465.g004

with the initiation of movements, while the Nogo or indirect pathway is associated with the inhibition of movements [39]. These two pathways include two separate populations of striatal neurons expressing different dopaminergic receptors [40]. The striatal Go neurons mainly express D1 receptors which are excited by dopamine, while the striatal Nogo neurons mainly express D2 receptors which are inhibited by dopamine [41]. Thus, dopaminergic activation signal controls the competition between these two pathways during action selection and promotes action initiation over inhibition.

Given the architecture of the basal ganglia, we hypothesise that this circuitry is well suited for the computation of the utility of actions in decision making. In particular, we note the similarities between the utility function in Eq (13) consisting of two terms that are differently affected by motivation (i.e. in the first term r is scaled by m, while in the second $-r^2/2$ is not), and the basal ganglia consisting of two main pathways that are differently modulated by dopamine. Accordingly, in the presented model the Go and Nogo neurons represent the estimates for these two terms and the dopaminergic activation signal encodes motivation and controls the relative influence of Go and Nogo neurons in an appropriate manner. Moreover, we propose that encoding r and $-r^2/2$ separately in the synaptic weights of Go and Nogo neurons, respectively, is beneficial as it allows for asymmetric weighting of the relative contributions of these terms by the dopaminergic activation signal. At first sight, this mapping may seem counter-intuitive because the second term of the utility equation is not modulated by motivation, while the Nogo neurons are suppressed by dopaminergic activity. However, in the next few paragraphs we show that the basal ganglia output encodes a quantity proportional to the utility when dopamine scales the relative effects of the two striatal populations on the output. Furthermore, separating the quantities that are scaled and are not scaled by motivation provides an additional benefit when considering a multi-dimensional space. As we highlight in

the discussion, some other factors that affect the utility of an action, e.g. related to effort, may be encoded in Nogo neurons but not in Go neurons.

Let us now consider the computation of the utility at the output of the basal ganglia. We refer to this output as the thalamic activity, denoted by T. T depends on the cortico-striatal weights of Go neurons (G) and Nogo neurons (N), and the dopaminergic activation signal denoted by D.

Although admittedly more complex, we can capture the signs of the influences of the dopaminergic activation signal, Go and Nogo neurons in a linear approximation [42]:

$$T = DG - (1 - D)N. \tag{14}$$

In the above equation, the contribution of Go neurons to the thalamic activity is described by the first term DG, reflecting facilitatory effect of dopamine on Go neurons. The inhibitory effect of Nogo neurons to the thalamic activity results in a negative contribution to the thalamic activity and is described by the second term $-(1 - D)N$. We assume that $D \in [0, 1]$, and that tonically active dopaminergic neurons give rise to a baseline level of the dopaminergic activation signal of $D = 0.5$ for which both striatal populations equally contribute to the thalamic activity.

In Eq (14) we make three assumptions about the contribution of Go and Nogo neurons to the thalamic activity and how they are modulated by the level of dopamine. For each of the assumptions we discuss how they relate to experimental data. 1) At baseline dopamine level $D = 0.5$, striatal Go neurons have a positive effect on the thalamic activity of the same magnitude as the negative effect striatal Nogo neurons have on the thalamic activity. This assumption is consistent with an observation that optogenetic activations of Go and Nogo neurons modulate the thalamic activity to the same extent but in opposite direction (see Figure 4 in [43]). 2) The dopaminergic activation signal increases the gain of Go neurons, and reduces the gain of Nogo neurons. This assumption is based on the observation of an increased slope of firing-input relationship of neurons expressing D1 receptors in the presence of dopamine [18], and decreased slope of neurons expressing D2 receptors in the presence of dopamine [19, 20]. 3) Changes in the level of dopamine affect the gain of Go and Nogo neurons to the same extent. This assumption differs from many models of dopaminergic modulation in which there is an unequal modulation of the Go and Nogo neurons based on their low and high receptor affinity, respectively, suggesting that Go neurons mainly respond phasic dopamine signals, whereas Nogo neurons mainly respond to tonic dopamine signals [44]. However, it is consistent with recent findings showing that Nogo receptors also responds to phasic changes in dopamine, something that is incompatible with the original models [45, 46]. Moreover, computational models that include both receptor affinity and receptor occupancy show that the effect of dopamine could be similar on Go and Nogo populations [47], which further supports our assumption.

We now show that the thalamic activity defined in Eq (14) is proportional to the utility of an action if G and N are fully learned and therefore provide correct estimates of the positive and negative terms in utility equation (Eq (13)), respectively ($G = r$ and $N = r^2/2$). Then, we can rewrite Eq (14) as:

$$T = Dr - (1 - D)r^2/2 \tag{15}$$

which can then be rewritten in the following way:

$$T = (1 - D)\left(\frac{D}{(1 - D)}r - r^2/2\right). \tag{16}$$

Comparing this to Eq (13), we observe that the thalamic activity is proportional to the utility ($T = (1 - D)U$) when the motivation is encoded by dopaminergic activation signal:

$$m = \frac{D}{1 - D}. \tag{17}$$

We can rewrite Eq (17) in the following way to express the level of dopamine for a given motivation:

$$D = \frac{m}{1 + m}. \tag{18}$$

In summary, when the striatal weights encode the positive and negative consequences and the dopaminergic activation signal is described by Eq (18), then the thalamic activity is proportional to the utility.

Let us now consider how this utility can be used to guide action selection. Computational models of action selection typically assume that all basal ganglia nuclei and thalamus include neurons selective for different actions [48]. Therefore, the activity of thalamic neurons selective for specific actions can be determined on the basis of their individual positive and negative consequences and the common dopaminergic activation signal. Given that the proportionality coefficient $(1 - D)$ in Eq (16) is the same for all actions, the utilities of different actions represented by thalamic activity are scaled by the same constant. This means that the most active thalamic neurons are the ones selective for the action with the highest utility, and hence this action may be chosen through competition. Furthermore, if we assume that actions are only selected when thalamic activity is above a threshold, then no action will be selected if all actions have insufficient utility. The utility of actions has to be sufficiently high to increase neural firing in the thalamus above the threshold and trigger action initiation. It is important to note that even though the proportionality coefficient $(1 - D)$ in Eq (16) approaches 0 when D approaches 1, this does not mean that the thalamic activity also approaches 0 and no action will be selected. When D approaches 1, the proportionality factor and the utility change concurrently, which means that the thalamic activity either stays the same or increases. More formally, the thalamic activity defined in Eq (14) is non-decreasing with respect to the dopaminergic activation signal, because $\mathrm{d}T/\mathrm{d}D = G + N$ which is non-negative (as synaptic weights G and N cannot be negative).

Furthermore, it may seem counter-intuitive that dopaminergic neurons modulate both Go and Nogo neurons, while the motivation m only scales the first term in the definition of the utility function of Eq (13). The benefit of such double modulation is that it allows for representing an unbounded range of the motivation, $m \in [0, \infty]$, while expressing the dopaminergic activation signal within a bounded range $D \in [0, 1]$. This is useful as biological neurons can only produce finite firing rates. Thus, if an animal is very hungry, setting the dopaminergic activation signal to a value close to a maximum value, denoted here by 1, allows for ignoring any negative consequences of an action. Consequently, the thalamic activity will be positive for any action associated with a positive reinforcement, facilitating the execution of these actions.

Models of learning. In the previous section, we showed that the basal ganglia network can estimate the utility once the striatal weights have acquired the appropriate values. In this section we address the question of how these values are learned. Earlier, we proposed a general framework for describing learning process assuming that the brain minimises a prediction error during this process and we redefined the prediction error as the difference between utility and expected utility. In the previous section we described a model in which the thalamic activity encodes a scaled version of the estimated utility ($\hat{U} = T/(1 - D)$ (see Eqs 16 and 17). This estimate of the utility can be substituted into Eq (7) giving the following the state-

dependent reward prediction error:

$$\delta = U - \frac{T}{(1-D)}. \tag{19}$$

In this RPE, U is the utility of an action (Eq (13)), and $T/(1 - D)$ is the expected utility of an action. In line with the general definition of Eq (7), the above equation shows that the RPE depends on a reinforcement and the expected reinforcement in a state-dependent manner, which is here instantiated in a specific model for estimating utility. Please note that at baseline levels of the dopaminergic activation signal ($D = 0.5$), the above expression for prediction error reduces to $\delta = U - (G - N)$ and such prediction error has been used previously [49].

We will now describe two models for learning the synaptic weights of G and N. The first model is a normative model, developed for the purpose of this learning, while the second model corresponds to a previously proposed model of striatal plasticity, which provides a more biologically realistic approximation of the first model. We have chosen this model as it is able to extract payoffs and costs. Other existing computational models fail to do so and thus we do not expect them to perform optimally (for reference see [42]).

Gradient model. The first model we used to describe learning of synaptic weights under changing conditions, directly minimises the error in prediction of the utility of action. It changes the weights proportionally to the gradient of the objective function:

$\Delta G = \alpha \partial F/\partial G$ and $\Delta N = \alpha \partial F/\partial N$, respectively. For the prediction error described in Eq (19), this gives us the following learning rules for G and N:

$$\Delta G = \alpha \delta \frac{D}{1-D}, \tag{20}$$

$$\Delta N = -\alpha \delta. \tag{21}$$

Synaptic weights of Go and Nogo neurons are updated using the dopaminergic teaching signal scaled by the learning rate constant α. The update rule for Go weights has an additional term involving the dopaminergic activation signal encoding the motivation as described in Eq (17). Only the update rule for G, but not for N, includes scaling by motivation, because in the definition of utility of Eq (13), the motivational level only scales the positive consequences of an action and not the negative.

Payoff-cost model. The second model has been previously proposed to describe how Go and Nogo neurons learn about payoffs and costs of actions. It has been shown to account for a variety of data ranging from properties of dopaminergic receptors on different striatal neurons to changes in risk preference when dopamine levels are low or high [49]. We expected this model to provide an approximation for the gradient model because it has been shown to be able to extract positive and negative consequences of actions. More specifically, if a reinforcement takes a positive value r_p half of the times and a negative value $-r_n$ the other half of times, then the Go weights converge to $G = r_p$ and Nogo weights to $N = r_n$, for certain parameters [42]. Therefore, we expected this learning model to be able to extract positive and negative terms of the utility in Eq (13) if motivation could vary between trials, so the positive term dominates utility on some trials while the negative term on other trials.

In our simulations we used the same update rules as previously described [42, 49], but we use a state-dependent prediction error (Eq (19)) to account for decision making under different physiological states.

$$\Delta G = \alpha f_{\epsilon}(\delta) - \lambda G, \tag{22}$$

$$\Delta N = \alpha f_{\epsilon}(-\delta) - \lambda N, \tag{23}$$

where

$$f_{\epsilon}(\delta) = \begin{cases} \delta, & \text{if} \quad \delta > 0, \\ \epsilon\delta, & \text{if} \quad \delta \leq 0. \end{cases}$$

The update rules in the above equations consist of two terms. The first term is the change depending on the dopaminergic teaching signal scaled by a learning rate constant α. It increases the weights of Go neurons when $\delta > 0$, and slightly decreases when $\delta < 0$, so that changes in the Go weights mostly depend on positive prediction errors. The constant ϵ controls the magnitude by which the weights are decreased, and takes values within a range $\epsilon \in [0, 1]$, meaning that for $\epsilon = 0$ the negative prediction error does not drive any decrease in Go weights. The details on the values of ϵ required for learning positive and negative consequences are provided in [42]. Nogo weights will be updated in a analogous way, but these changes mostly depend on negative prediction errors. The second term in the update rules is a decay term, scaled by a decay rate constant λ. This term is necessary to ensure that the synaptic weights stop growing when they are sufficiently high and allows weights to adapt more rapidly when conditions change. In case an updated weight becomes negative, it is set to zero.

Simulations of learning. In this section, we investigate under what conditions the learning rules described above can yield synaptic weights of Go and Nogo neurons that allow for the estimation of utility. Recall that the network will correctly estimate the utility, if $G = r$ and $N = r^2/2$.

In the simulations we make the following assumptions: 1) The simulated animal knows its motivational level m, which influences both dopaminergic signals accordingly (Eqs (18) and (19)). 2) The simulated animal computes the utility of obtained reinforcement as a change in the desirability of the physiological state. As described above, the desirability depends on the objective value of the reinforcement r and the current motivational state m according to Eq (13), which was used to compute the reward prediction error according to Eq (19).

We simulated scenarios in which the simulated animal repeatedly chooses a single action and experiences a particular reinforcement r under different levels of motivation $m \in \{m_{\text{low}}, m_{\text{baseline}}, m_{\text{high}}\}$. Note that the $m_{\text{low}} = 0$ correspond to a dopaminergic activation signal of $D = 0$, $m_{\text{baseline}} = 1$ gives a dopaminergic activation signal of $D = 0.5$, which means that Go and Nogo neurons are equally weighted, and $m_{\text{high}} = 2$ corresponds to a dopaminergic activation signal above baseline levels.

We first simulated a condition in which the motivation changed on each trial, and took a randomly chosen value from a set $\{m_{\text{low}}, m_{\text{baseline}}, m_{\text{high}}\}$ (Fig 5A). The gradient model was able to learn the desired values of Go and Nogo weights as Go weights converged to r, while Nogo weights converged to $r^2/2$, which allowed the network to correctly estimate the utility. Although the subjective reinforcing value changed as a function of physiological state, the model was able to learn the actual reinforcement of an action. Encoding of such objective estimates allows the agent to dynamically modulate behaviour based on metabolic reserves. In contrast, the payoff-cost model converged to lower weights than desired. Although it learned the synaptic weights based on the state-dependent prediction error, the weight decay present in the model resulted in a lower asymptotic value.

To test robustness of the learning rules and because the motivational state is fixed during the experimental paradigms simulated in this paper, we also simulated conditions in which the motivation was kept constant (Fig 5B–5D). In these cases both leaning rules converged to very

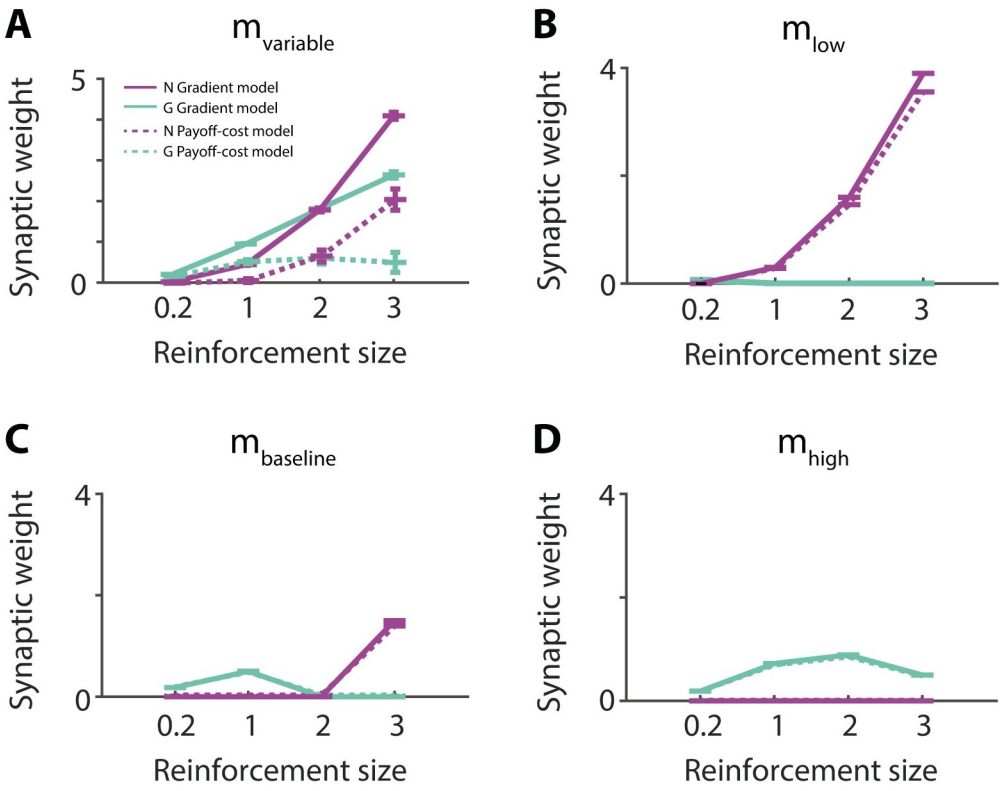

Fig 5. Learning Go and Nogo weights for different reinforcements and different levels of motivation. Performance of models under variable (A), low (B), baseline (C), high (D) levels motivation. Simulations were performed using the state-dependent prediction error (Eq (19)). Solid lines show simulations of the gradient model using the plasticity rules described in Eqs (20) and (21). Dashed lines show simulations of the payoff-cost model using the plasticity rules described in Eqs (22) and (23). Black lines correspond to Nogo neurons and grey lines to Go neurons. Each simulation had 150 trials and was repeated 100 times.

https://doi.org/10.1371/journal.pcbi.1007465.g005

similar values of synaptic weights: low levels of motivation emphasised negative consequences and therefore facilitated Nogo learning (Fig 5B), while high levels of motivation emphasised positive consequences and therefore facilitated Go learning (Fig 5D). This shows that when the motivation is equal to the reward size ($m = r$), meaning that this action brings the subject to exactly the desired state S^*, the synaptic weights of Go neurons are higher than Nogo neurons $G > N$. Whenever the action exceeds the desired state, $r > m$, the Nogo weights increase as this action is not favourable $N > G$.

In summary, the simulations indicate that for the models to learn appropriate values of synaptic weights, the reinforcements need to be experienced under varying levels of motivation. In this case, the gradient model provides a precise estimation, while the payoff-cost model provides an approximation of the utility. In cases when the motivational state is fixed during training, both models learn very similar values of the weights.

The basal ganglia architecture allows for efficient learning. In the previous sections we presented models and analysed how utilities can be computed and learned in the basal ganglia network. One could ask, why would the brain employ such complicated mechanisms if a simple model could give you the same results? In particular, one could consider a standard Q-learning model, in which the state is augmented by motivation. Such model would also be able to learn to estimate the utility. However, such a model does not incorporate any prior knowledge about the form of the utility function and its dependence on motivation. By contrast, the

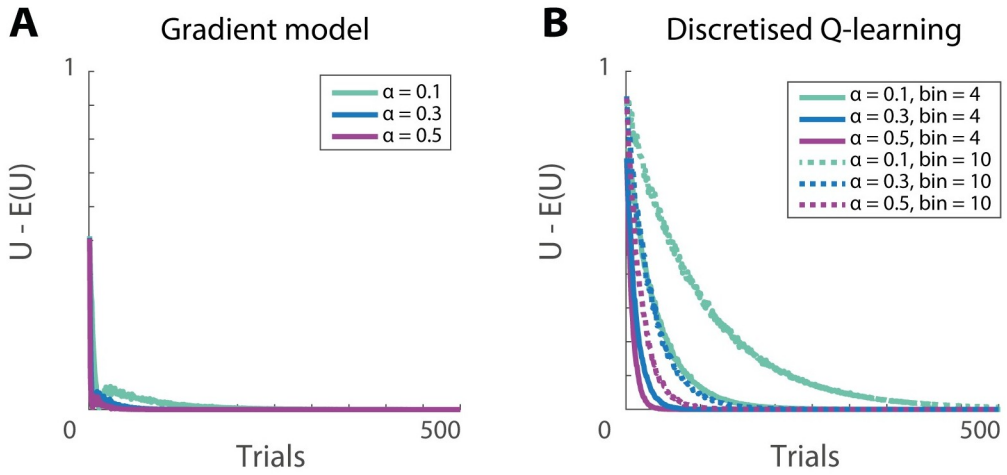

Fig 6. Reward prediction error as a function of learning iteration. A) The gradient model uses the state-dependent prediction error (Eq (19)) and the plasticity rules described in Eqs (20) and (21). B) Discretised Q-learning model. Motivational values were randomly chosen on each trial from a uniform distribution between 0 and 2. For Q-learning, motivational values were binned in either 4 or 10 bins. The *y*-axis corresponds to reward prediction error equal to the difference between the estimated and expected utility.

https://doi.org/10.1371/journal.pcbi.1007465.g006

model grounded in basal ganglia architecture, assumes a particular form of the utility function to be learned. In machine learning, such prior assumptions are known as 'inductive bias', and they facilitate learning [50].

We now illustrate that thanks to the correct inductive bias, the gradient model learns to estimate the utility faster than standard Q-learning, which does not make any prior assumptions about the form of the utility function. In our implementation of Q-learning, the range of values the motivation can take was divided into a number of bins, and the model estimated the utility for each bin. In the simulations on each trial reinforcement $r = 1$ was received and its utility was computed using Eq (13), which relied on the current motivation. The Q-value for the current motivation bin was updated by: $\Delta Q_m = \alpha(U - Q_m)$. In Fig 6 we compare a Q-learning approach, in which the motivational state was discretised, with the gradient model in our framework, which does not require discretisation of the motivational state. As can be seen in in Fig 6, both models are able to approximate the utility well. However, Q-learning takes significantly more trials to do so. Moreover, the more bins are used for the discretisation, the slower the learning occurs.

Relationship to experimental data

We already demonstrated how a model employing an approximation of utility can explain data on the effect of physiological state on dopaminergic responses. In this section, we demonstrate how we can use models grounded in basal ganglia architecture to describe these dopaminergic responses and goal-directed action selection in different experimental paradigms.

We first show that the new, more complex and biological relevant learning rules can also be used to explain the data by Cone et al. [26]. In these simulations, the dopaminergic teaching signal at the time of the CS took on the value of the expected utility ($T/(1 - D)$) and at the time of the US represented the reward prediction error described by (Eq (19)). Simulated values of the dopaminergic teaching signal (Fig 7) show similar behaviour to the experimental data by Cone et al. [26]. Both the gradient and the payoff-cost model produce a similar dopaminergic teaching signal. This could be expected from simulations in the previous sections, which

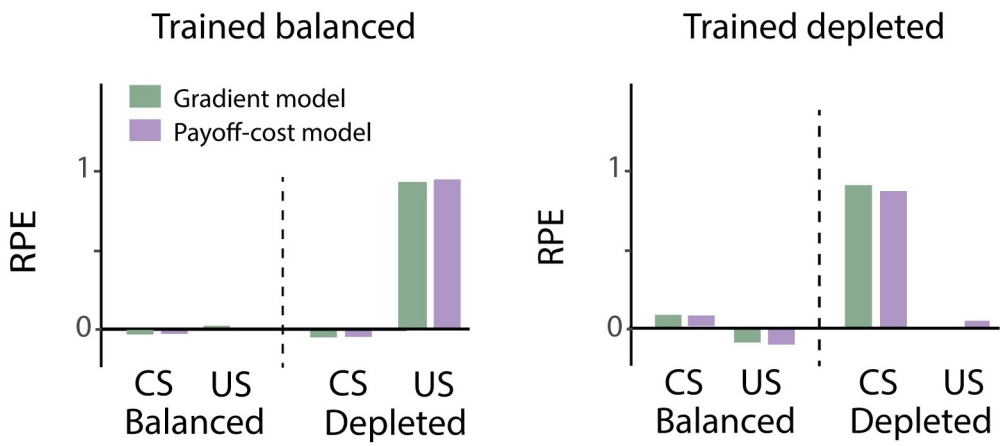

Fig 7. Simulated dopaminergic teaching signal in the paradigm of Cone et al. [26] according to models grounded in basal ganglia architecture. For all simulations, the state-dependent prediction error (Eq (19)) was used. The gradient model, depicted in grey, uses the plasticity rules described in Eqs (20) and (21). Payoff-cost model, depicted in black, uses the plasticity rules described in Eqs (22) and (23). Left and right panels show the data tested in the balanced state or depleted state, respectively. CS = conditioned stimulus, US = unconditioned stimulus, RPE = reward prediction error. Each simulation consisted of 50 training trials, 1 test trial and was repeated 5 times, similar to the number of animals in each group.

https://doi.org/10.1371/journal.pcbi.1007465.g007

showed that both models converge to similar weights if the motivation is kept constant during training.

Influence of physiological state on action selection. In the presented framework natural appetites, such as hunger or thirst, can drive action selection into the direction of the relevant reinforcement. Generally speaking, most food items are considered appetitive even when an animal is in the near-optimal state. Nevertheless, overconsumption could have negative consequence as you can experience discomfort after eating too much. Therefore some of these negative consequences might have to be accounted for as well. As discussed above, a good example of a natural appetite that can be both appetitive and aversive dependent on the physiological state of the animal is salt appetite. Salt is considered very aversive or appetitive when the sodium physiology is balanced or depleted, respectively. Accordingly, rats reduce their intake of sodium or salt-associated instrumental responding when balanced and vice versa when depleted [27, 51]. This even occurs when animals have never experienced the deprived state before and have not had the chance to relearn the incentive value of a salt reinforcement under a high motivational state [51]. This example fits very well with the incentive salience theory which states that the learned association can be dynamically modulated by the physiological state of the animal. Modulation of incentive salience adaptively guides motivated behaviour to appropriate reinforcements.

To demonstrate that the simple utility function (Eq (13)) proposed in this paper can account for the transition of aversive to appetitive reinforcements, and vice versa, in action selection, we use the study by Berridge and Schulkin [27]. In this study, animals learned the value of two different conditioned stimuli, one associated with salt intake (CS+) and one with fructose intake (CS-). The animals were trained in a balanced state of sodium. Once the appropriate associations had been learned, the animals were tested in a sodium balanced and sodium depleted state. As can be seen in Fig 8A the intake of the CS+ was significantly increased in the sodium depleted state compared to the balanced state and the CS- intake. If we assume that positive and negative consequences are encoded by the Go or Nogo pathway, respectively, the synaptic weights of these pathways will acquire positive or negative values depending on the

situation. Again, the dopaminergic activation signal can control to what extent these positive and negative consequences affect the basal ganglia output as Go and Nogo neurons are modulated in an opposing manner.

Once the appropriate associations between the conditioned stimuli and the outcomes are acquired, the outcomes can be dynamically modulated by the relevant state only (i.e. the level of sodium depletion). The fact that the responses to the CS- are unaffected by the physiological state of sodium suggests that salt and fructose are modulated by separate appetitive systems and that the physiological state of the animal modulates the intake proportional to the deprivational level of the animal. The phenomenon that different reward types act on different appetitive system has been also observed by other experimental studies [25].

In our simulation, we assumed that the synaptic weights for Go and Nogo neurons were learned in a near-balanced state of sodium since the animals had never experienced a sodium depleted state before. During training, the motivation was low ($m = 0.2$), resulting in low level of dopaminergic activation signal following Eq (18). During the testing phase, the motivation for the CS+ was low ($m = 0.2$) for sodium in the near-balanced state and high ($m = 2$) for the sodium depleted state. Given that experimental data suggests that multiple appetitive systems may be involved we used separate motivational signals for the CS+ and CS-. Therefore, the motivation for the CS- were kept low ($m = 0.1$), but were non-negative, for both sodium near-balanced and sodium depleted states since fructose has no effect on the physiological state of sodium and we assumed that the animals were not deprived of other nutrients. The thalamic activity was computed using Eq (14), and additional Gaussian noise was added to allow exploration. Actions were made when the thalamic activity was positive, otherwise no action was made. The model received a reinforcement of $r = 0.5$ for each action made and the utility was computed using Eq (13). During training, Go and Nogo weights were updated using the update equations presented above for the different models. For the testing phase, the Go and Nogo values were kept constant based on the learned values and were not allowed to be (re-) learned. Again, the thalamic activity was computed and actions were taken when this was positive. Please note that the main difference between near-balanced and depleted states, is the level of dopaminergic activation signal. As can be seen in Fig 8B both models show dynamic scaling of the CS+ dependent on the relevant motivational state similar to the experimental data in Fig 8A.

State-dependent valuation. There is a number of experimental studies that have investigated the influence of physiological state at the time of learning on the preference during subsequent encounters (e.g. [28, 35, 36, 52]). In the study by Aw et al. [28], animals were trained in both a near-balanced and deprived state. One action was associated with food in the near-balanced state and another action was associated with food in the deprived state. Animals were tested in both states. In both cases, animals preferred the action associated with the deprived state during learning and the proportions of trials with these actions is above chance level (Fig 9A). These results resemble the data on dopaminergic responses (Fig 1), which also demonstrated higher response to reward-predicting stimuli (CS) that had been experienced in a depleted state. In this section we show that such preferences can also be produced by the proposed models.

We simulated learning of the synaptic weights of Go and Nogo neurons when the motivation was high (i.e. hungry) and when the motivation was low (i.e. sated). In the experiment by Aw and colleagues, the training phase consisted of forced choice trials in which the reinforcement was only available in one arm of a Y-maze while the other arm was blocked. For example, the left arm was associated with a food reinforcement during hunger and the right arm was associated with a food reinforcement during the sated condition. In the experiment, 11 animals were used, which were trained for 65 trials on average to reach the required performance.

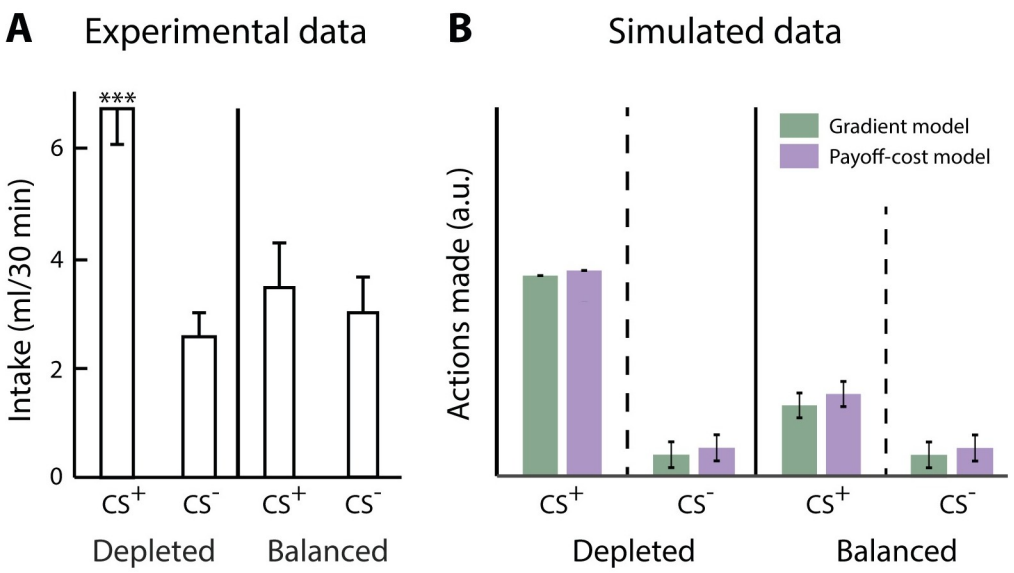

Fig 8. Salt appetite: increased approach behaviour for salt when salt-deprived. A) Re-plotted data of Berridge and Schulkin [27] showing that the intake of sodium (CS+), but not fructose (CS-), is increased when salt-deprived. B) Simulated data of number of actions made using the state-dependent prediction error (Eq (19)). The gradient model, depicted in grey, uses the plasticity rules described in Eqs (20) and (21). The payoff-cost model, depicted in black, uses the plasticity rules described in Eqs (22) and (23). Within the graph, the left and right halves show the responses of animals tested in depleted and balanced states, respectively. CS+ = relevant conditioned stimulus for sodium, CS- = irrelevant conditioned stimulus for fructose.

https://doi.org/10.1371/journal.pcbi.1007465.g008

In line with this, our simulations were repeated 11 times, and in each iteration we trained the model for varying trial numbers with a mean of 65. Motivation was set to $m = 2$ for hungry and $m = 0.2$ for sated. The dopaminergic activation signal was fixed to values that correspond to the motivation described by Eq (18). For each correct action, the model received a

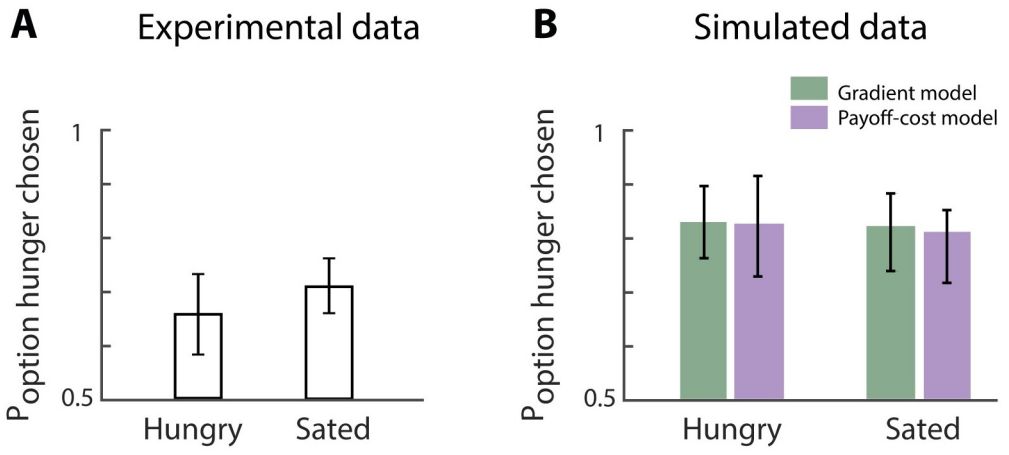

Fig 9. Valuation based on state at training, not testing, predicts preference in choice behaviour. A) Re-plotted experimental data by Aw et al. [28] showing that the option associated with hunger during training is always preferred regardless of the state at testing. B) Simulated data using the state-dependent prediction error (Eq (19)). Gradient model, depicted in grey, uses the plasticity rules described in Eqs (20) and (21). Payoff-cost model, depicted in black, uses the plasticity rules described in Eqs (22) and (23). Hungry and sated refer to the physiological state at testing. The proportion of actions for the arm associated with the hunger condition during training is depicted on the y-axis. For the simulated data this is counted as the number of times the thalamic activity was positive during the test for either the option associated with hunger or sated condition during training.

https://doi.org/10.1371/journal.pcbi.1007465.g009

reinforcement of $r = 0.2$ and the utility was calculated using Eq (13). At the start of each simulated forced trial, the model computed the thalamic activity (using Eq (14)) of the available action and some independent noise was added. The thalamic activity for the unavailable action was zero. The action with the highest positive thalamic activity was chosen. If the thalamic activity of all actions was negative, no action was made and the reinforcement was zero. Each time an action was made the synaptic weights of Go and Nogo neurons were updated using the state-dependent reward prediction error and the update rules described in the section Models of learning. The learning rate for all of these models was set to $\alpha = 0.1$. Once learning was completed, the synaptic weights were fixed and were not updated during the testing phase.

During the testing phase, both arms were available and the animals could freely choose an arm to obtain a reinforcement in. All 11 animals were tested for 24 trials. The simulated tests were run for 24 trials for both conditions and repeated 11 times using the individual learned Go and Nogo weights for the sated and hungry condition. Again, the model computed the thalamic activity for both options simultaneously (in parallel) plus some independent noise. The action with the highest thalamic activity was chosen. The proportion of actions associated with the hungry option are depicted in Fig 9B. This experiment was simulated for both physiological states during testing phase. The proportion of actions for the arm associated with hunger were calculated for both states. Both the experimental and simulated data show that the animals chose the action associated with the hungry state more often, regardless of the current state.

To gain some intuition for why the models preferred the option that was associated with hungry state during training, let us look back at the simulations presented in Fig 5B–5D. These show that when the models were trained with a fixed motivation, Go weights took higher values when the motivation was high during training, and Nogo weights were larger when motivation was low. Analogously, in the simulations of the study of Aw and colleagues, the Go weights took larger values for the option associated with hunger during training (Fig 10), and this option was therefore preferred during testing. These biases arise in the models when they are unable to experience multiple levels of motivation during training. Only after

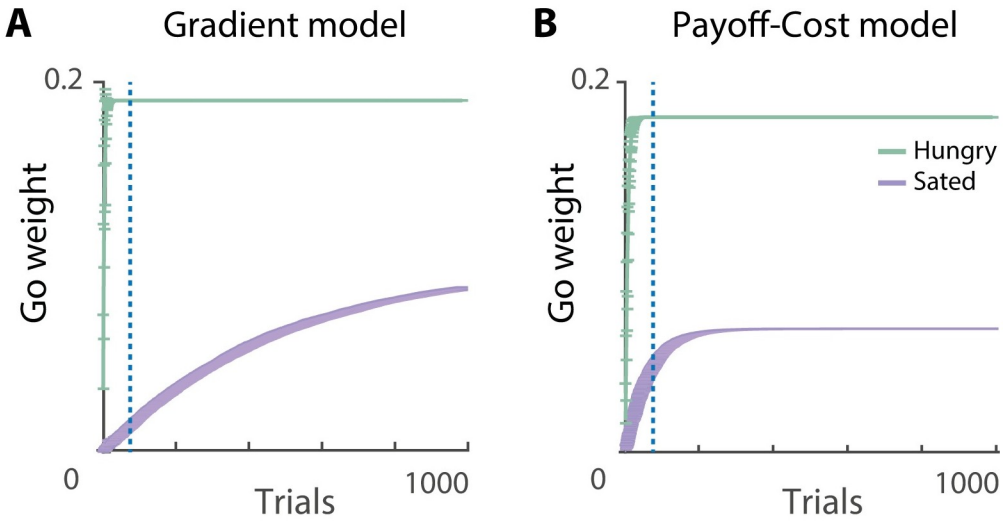

Fig 10. Learning Go weights as function of physiological state. Go weights were updated using Eqs (20) and (22) for the gradient (A) and payoff-cost model (B), respectively. Learning at high motivation is depicted in green and learning at low motivation is depicted in purple. Number of trials used for the simulation was 1000. Go weights were initialised at 0.1.

https://doi.org/10.1371/journal.pcbi.1007465.g010

training in multiple physiological states the utility can be flexibly estimated in different physiological states.

Comparison to an existing incentive salience model

In this section we formally compare the proposed framework with existing models for incentive salience [38, 53]. Zhang et al., [38] developed mathematical models that describe how the values of reinforcement are modulated by changes in a physiological state. In these models the learned value of a reinforcement, which we denote by R, takes positive values for appetitive reinforcements and negative values for aversive reinforcements. A change in the physiological state of the animals from training to testing is captured by a gating parameter κ, which is always non-negative, $\kappa \in [0, \infty)$. An up-shift in physiological state (i.e. when an animal becomes more deprived) is represented by $\kappa > 1$, while a down-shift in physiological state (i.e. when an animal becomes less deprived) is represented by $\kappa < 1$. Up-shifts increase the reward value whereas down-shifts decrease the reward value. Two mechanisms are considered through which a shift in physiological state, κ, modulates the reinforcement, R [38]. According to the first, multiplicative mechanism, a subjective utility of reinforcement is equal to $R\kappa$. This multiplicative mechanism is similar to the first order approximation of the utility (Eq (4)). However, as the model of Zhang et al. assumes that $\kappa > 0$ for all states of the animal, it struggles to explain a phenomenon such as salt appetite where aversive reinforcements can become appetitive depending on the physiological state of the animal, because multiplying a negative reinforcement with a positive factor does not change the sign of the reward value. To account for this phenomenon, Zhang et al. developed a second, additive mechanism, where the subjective utility is equal to:

$$U(CS) = R + \log(\kappa) \tag{24}$$

For sufficiently large or small κ, this mechanism is able to change the polarity of the utility. As the models of Zhang et al. do not describe learning, it is not possible to compare them explicitly in the simulations of the learning tasks presented in this paper. Furthermore, these models do not define reward prediction error, so it also not possible to compare their predictions with dopaminergic responses to rewarding stimuli. Nevertheless, we compare the utilities computed according to the additive model with the dopaminergic responses to conditioned stimuli in the study of Cone et al., as such responses are thought to reflect the learned values of the stimuli. Fig 11 shows utilities for the four experimental conditions computed according to Eq (24). The value of reinforcement parameter was set to $R = -1$ in the balanced state as salt is aversive in that state, and to $R = 1$ in the depleted state because it is then apetitive. The incentive salience gating parameter was set to $\kappa = 1$ when there was no change between the trained state and the tested state. It was set to $\kappa > 1$ when there is an increase in the reward value in the tested state compared to reward value in the trained state and to $\kappa < 1$ when there was a devaluation of the reward value of the tested state compared to reward value in the trained state. The particular values of κ were chosen in such a way that make the utilities qualitatively resemble the dopaminergic responses observed by Cone et al.

Although the model can account for the dependence of dopaminergic responses on the testing state when animals were trained in the depleted state, the values of conditioned stimuli in the model do not correspond to the dopaminergic responses Cone et al. observed when animals were trained in the balanced state. The dopaminergic responses to the CS are both close to baseline for animals trained in the balanced state regardless of the state at testing. The additive model does not produce similar responses because different values of $\log \kappa$ are added in the two testing states. Please note that this qualitative difference between the utilities computed

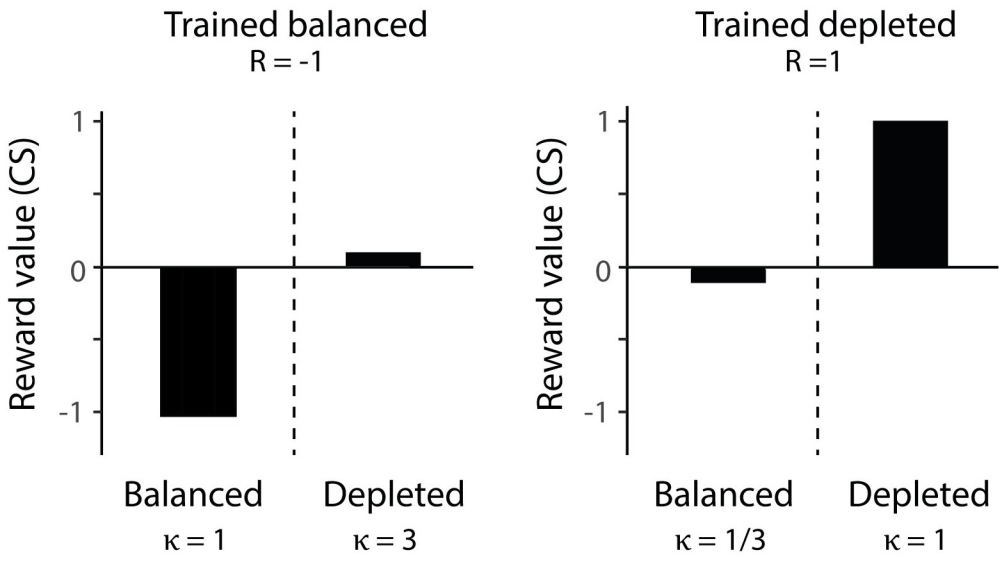

Fig 11. The additive incentive salience model by Zhang et al. [38] predicts different patterns in dopaminergic teaching signals. Utility of conditioned stimuli in a study by Cone et al. [26] was computed according to Eq (24). The reward value was set to $R = -1$ for the animals trained in the balanced state and $R = 1$ for animals trained in the depleted state. The gating parameter $\kappa = 1$ was when the training state is equal to the testing state. An up-shift in state is represented by $\kappa = 3$ and a down-shift in state is represented by $\kappa = 1/3$.

https://doi.org/10.1371/journal.pcbi.1007465.g011

by the additive model and the dopaminergic responses following training in the balanced state remain present, irrespective of the choice of the value of R: Even if we set $R = 0$ to capture that no learning occurred in the balanced state, the utilities of the stimuli remain different (both shifted upwards), as different values of log κ are added.

In conclusion, the most obvious difference between our model and the models by Zhang et al. is the form of the utility function. We developed one mechanism to describe all types of behaviours, whereas Zhang et al. rely on two mechanisms to account for different scenarios. Furthermore, our model is able to describe the learning process by which the estimates of a reward value can be learned. It is important to note that Zhang et al. uses a "gating" parameter which describes the direction and the size of the shift between the previous state and the new state, whereas the incentive salience in our models is captured by the parameter for motivation m which describes the current state of an animal. The transition itself and the effect it has on action selection and learning is captured by the state-dependent reward prediction error and the thalamic activity.

Discussion

In this paper, we have presented a novel framework for action selection under motivational control of internal physiological factors. The novel contribution of this paper is that the framework brings together models of direct and indirect pathways of the basal ganglia with the incentive learning theory, and proposes mechanistically how the physiological state can affect learning and valuation in the basal ganglia circuit. We proposed two models that learn about positive and negative aspects of actions utilising a prediction error that is influenced by the current physiological state. In this section, we will discuss the experimental predictions, the relationship to experimental data and other computational models and other implications.

Experimental predictions

In this section, we outline the predictions the models make. In our model we hypothesise that the synaptic weights of Go neurons and Nogo neurons are equal to r and $r^2/2$, respectively. Although a relationship between reward and Go/Nogo neuron activity has been demonstrated [54], the effects are heterogeneous which makes an exact mapping of G and N difficult. The mapping chosen in this paper may seem arbitrary, however, from a biological point of view it captures the effects of the sum of Go and Nogo neurons with respect to reward, and the relationship between the thalamic activity and utility with respect to the motivation and the reward that matches the experimental data. Alternative mappings with the same overall effect may be plausible as well and will need to be tested in experimental studies as suggested below. The neural implementation of the framework assumes that Nogo neurons prevent selecting actions with large reinforcements when the motivation is low. Thus, it predicts that pharmacological manipulations of striatal Nogo neurons through D2 agonist (or antagonist) should increase (or decrease) the animal's tendency to consume large portions of food or other reinforcers to a larger extent when it is close to satiation, than when it is deprived.

The neural implementation of the framework also assumes that the activity in Go and Nogo pathways is modulated by the dopaminergic activation signal, which depends on motivation. This assumption could be tested by recording activity of Go and Nogo neurons, for example using photometry, while an animal decides whether to consume a reinforcement. The framework predicts that deprivation should scale up responses of Go neurons, and scale down the response of Nogo neurons.

As showed in Fig 5A, the framework predicts that the synaptic weights of Go and Nogo neurons converge to different values depending on the reinforcement magnitude. These predictions can be tested in an experiment equivalent to the simulation in Fig 5A in which mice learn that different cues predict different reinforcement sizes, and experience each cue in variety of motivational states. The weights of the Go and Nogo neurons are likely to be reflected by their neural activity (e.g. measured with photometry) while an animal evaluates a cue at baseline motivation level. We expect both populations to have higher activity for cues predicting higher reinforcements, and additionally, the gradient model predicts that the Go and NoGo neurons should scale their activity with reinforcement magnitude linearly and quadratically, respectively.

Relationship to experimental data and implications

The proposed framework can account for decision making and learning as a function of physiological state, as shown by the simulations of the data by Cone et al., Berridge and Schulkin and Aw et al. More specifically, we proposed that learning occurs based on the difference between the utility and expected utility of an action. This is in line with results from a study in monkeys that also suggested that dopaminergic responses reflects a difference in utility of obtained reward and expected utility [31]. That study focused on a complementary aspect of subjective valuation of reward, namely that the utility of different volumes of reward is not equal to the objective volume, but rather to its nonlinear function. In this paper we additionally point out that the utility of rewards depends on the physiological state in which they are received.

Furthermore, we know from literature that low levels of dopamine, as seen in Parkinson's disease patients, drive learning from errors, whereas normal/high levels dopamine emphasise positive consequences [3, 5, 55]. We also know from human imaging studies that hungry people show an increased BOLD response to high calorie food items, whereas sated people show an increased BOLD response to low calorie food items [56]. In our simulations we observe this

as well: a low dopaminergic activation signal favours small rewards and emphasises the negative consequences of actions encoded in the synaptic weights of Nogo neurons. In contrast, a high dopaminergic activation signal favours large rewards and emphasises the positive consequences of actions encoded in the synaptic weights of Go neurons. However, there are a couple considerations that have to be made with respect to the dopaminergic signal in our simulations. First, we assume that striatal neurons can read out both motivational and teaching signals encoded by dopaminergic neurons [22]. In our theory, we describe two roles of dopamine neurons, namely activation and teaching signal, however, we do not provide a solution to how these different signals are accessed. The function of dopamine neurons has been under a current debate and its complexity is not well understood [23]. We will leave the details of the mechanisms by which they can be distinguished to future work. We assume that the models, particularly the gradient model, has access to multiple dopaminergic signals simultaneously. Although we recognise that this is a simplified concept of what might be happening in the brain, it still provides us with new insights in how these different functions affect aspects of decision making. Further research is necessary to describe the complexity of dopamine neurons in decision making.

Second, it needs to be clarified how the prediction error in our model can be computed. According to Eq (19), the dopaminergic teaching signal depends on the value of dopaminergic activation signal (D) which scales thalamic input. One possible way in which such prediction error could be computed, without the dependence of one dopaminergic signal on another, could involve scaling the synaptic input from thalamic axons directly by an input encoding motivational state (m). Indeed, dopamine neurons receive inputs from the hypothalamus which is under the control of peripheral hormones signalling energy reserves and are also modulated directly by these hormones. More specifically, the orexigenic hormone ghrelin increases dopamine release [57], whereas the anorexigenic hormone leptin decreases dopamine release from the midbrain [58].

Third, in this paper we have focused primarily on one dimension, namely nutrient deprivation. However, experimental data suggests that reinforcements are scaled selectively by their physiological needs [25]. A nutrient specific deprivation alters goal-directed behaviour towards the relevant reinforcement, but not the irrelevant one. In contrast, other physiological factors, such as fatigue, may scale only the negative, but not the positive consequences. This hypothesis is supported by data showing that muscular fatigue alters dopamine levels [59]. Together this suggests that the utility of an action is most likely the sum of all the positive and negative consequences with respect to their physiological needs or other external factors. Therefore, extending the current theory to multiple dimensions is an important direction of future work. In such an extended model, an action which changes the state of multiple physiological dimensions, e.g. hunger, thirst and fatigue, would need to be represented by multiple populations of Go and Nogo neurons. In this example, the value of food and drink reinforcement, r_f and r_d respectively, would need to be represented by separate populations of Go and Nogo neurons that are modulated by different populations of dopaminergic neurons encoding information about hunger and thirst, while the effort would need to be encoded by a population of Nogo neurons modulated by a fatigue signal. It would be interesting to investigate if the required number of neurons could be somehow reduced by grouping terms in the utility function scaled in a similar way (e.g. $-r_f^2/2 - r_d^2/2$, which are not scaled by any factor in this example). As explained in the Results section, such factors are represented in the model by the Nogo neurons. Future experimental work on the diversity of how individual dopaminergic neurons are modulated by different physiological states would be very valuable in constraining such an extended model.

For simplicity, we used the same framework to model experiments involving both classical and operant conditioning. However, these learning processes are different as operant conditioning requires an action to be performed, while classical conditioning does not. To capture the difference between these paradigms, reinforcement learning models assume that operant conditioning involves learning action values, while classical conditioning involves learning the values of states, and analysis of neuroimaging data with such models suggests that these processes rely on different subregions of the striatum [60]. It would be interesting to investigate in more detail how the circuits in the ventral striatum could evaluate the utility of states.

Although the current study does not focus on risk preference, there is some evidence for the existence of a link between risk preference and physiological state [33, 61]. Particularly in the payoff-cost model, the dopaminergic activation signal controls the tendency to take risky actions [49], thereby predicting that motivational states such as hunger can increase risk seeking behaviour. The above mentioned studies show that changes in metabolic state systematically alter economic decision making for water, food and money and correlate with hormone levels that indicate the current nutrient reserve. Individuals became more risk-averse when sated whereas people became more risk-seeking when food deprived.

The current study also provides insights into the mechanistic underpinnings of overeating and obesity. Imaging studies using positron emission tomography showed an important involvement of dopamine in normal and pathological food intake in humans. In comparison to healthy controls, pathologically obese subjects show reduced availability of striatal D2 receptors that were inversely associated with the weight of the subject [62–64]. Our theory suggests that the ability to restrain from taking actions and learn from negative consequences of actions such as overeating may be diminished when D2 receptors are activated to a lesser extent. The involvement of the DA system in reward and reinforcement suggests that low engagement of Nogo neurons in obese subjects predisposes them to excessive use of food.

Relationship to other computational models

In the Results section we have already briefly compared our model to existing computational models of incentive salience by addressing some similarities and differences and we pointed out how our models could overcome some of the challenges the other models struggle with. Here, we relate our framework to several other theories.

One of these theories was proposed by Keramati and Gutkin [29] who developed a theory that also extended the reinforcement learning theory to incorporate physiological state. They defined a 'homeostatic space' as a multidimensional metric space in which each dimension represents a physiologically-regulated variable. At each time point the physiological state of an animal can be represented as a point in this space. They also define motivation (to which they refer to as 'drive') as the distance between the current internal state and the desired set point. We extended this theory to include how the brain computes the modulation of learned values by physiology.

In the motivation for the existence of the desired physiological state, Keramati and Gutkin [29] referred to active inference theory [65]. Our framework also shares a conceptual similarity with this theory, in that both action selection and learning can be viewed as the minimisation of surprise. To make this link clearer, let us provide a probabilistic interpretation for action selection and learning processes in our framework. This interpretation is inspired by a model of homeostatic control [66]. It assumes that the animal has a prior expectation $P(S)$ of what the physiological state S should be, which is encoded by a normal distribution with mean equal to the desired state S^*. That model assumes that animals have an estimate of their current bodily state S (interoception). It proposes that animals avoid states S that are unlikely according to

the prior distribution with mean S^* (thus they minimise their "interoceptive surprise"), and they wish to find themselves in the states S with high prior probability $P(S)$. Following these assumptions, we can define the desirability of the state as $Y(S) = \ln P(S)$. If we assume for simplicity that $P(S)$ has unit variance and ignore an additive constant, we obtain our definition of a desirability of a state in Eq (2). In our framework, actions are chosen to minimise the surprise of ending up in a new physiological state. The closer this state is to the desired state the more likely it will be and the smaller the surprise. Furthermore, motivation itself could be viewed as an error in the prediction of the physiological state.

Similar to action selection, animals update the parameters of their internal model (e.g. V, G, N) during learning in order to be less surprised by the outcome of the chosen action. To describe this more formally, let us assume that the animal expects the utility to be normally distributed with mean $\hat{U}$ and variance 1 (for simplicity). Furthermore, assume that during learning the animal minimises the surprise about the observed utility of action U. Therefore, we can define the negative of this surprise as $F = \ln P(U)$. This objective function is equal (ignoring a constants) to our objective function defined in Eq (9). Thus in summary, similar to the active inference framework, both action selection and learning could be viewed as processes of minimising prediction errors.

There are many computational models developed for action selection in either very abstract or more biological relevant ways. One of the leading models in describing how dopamine controls the competition between the Go and Nogo pathways during action selection is the Opponent Actor Learning (OpAL) model. This model hypothesises that the Go and Nogo neurons encode the positive and negative consequences of actions respectively [67] and high dopamine levels excite Go neurons and low levels of dopamine releases the inhibition of Nogo neurons. Moreover, existing neurocomputational theories describe how experience modifies striatal plasticity and excitability of the Go and Nogo neurons as a function of reward prediction errors [13, 48, 49, 67]. In line with action selection models of the basal ganglia we assumed that the Go neurons encode positive consequences and Nogo neurons encode negative consequences and that dopaminergic activation signal controls the balance between these neurons. We extended these concepts by combining them with incentive salience theory.

Other models of action selection suggest that the dopaminergic activation signal is associated with an increase in the vigour of actions [10]. In the study by Niv et al. [10] the same assumption is held that the utility of the reinforcement is dependent on the deprivational level, however, they do not provide a mechanism for how these utilities are computed and are therefore set them arbitrarily. Moreover, they rely on average reward reinforcement learning techniques which reveal an optimal policy that leads to an average reward rate per time unit. Following this line of thinking, actions with higher utility (i.e. actions taken in a deprived state) cause higher response rates as the opportunity cost of time increases. Although our model does not describe vigour or response times, it could be related to these output statistics thanks to recent work investigating the relationship between activity of a basal ganglia model and the parameters of a diffusion model of response times in a two alternative choice task [68]. This study showed that a drift parameter of a diffusion model is related to the difference in the activation of Go neurons selective for the two options, while the threshold is related to the total activity of Nogo neurons. In our framework motivation scales linearly with Go neurons for both options; it enhances the difference in their activity. Based on the data by Dunovan et al. [68] motivation is expected to increase the drift rate and reduce the threshold leading to faster responding.

Our framework considers for simplicity that all physiological dimensions (e.g. hunger, salt level, body temperature) have unique values with a maximum desirability, and the

desirability is lower for both smaller and higher values along the physiological dimension. Although this assumption is realistic for some dimensions, it may not be realistic for the resources animals store outside their bodies. Indeed, according to the classical economic utility theory, humans always wish to have more monetary resources. Although our framework also assumes diminishing utility of larger reinforcements, it differs from the classical theory in that the utility is a non-monotonic function of reinforcement. We also assume that all the resources are directly consumed, which does not allow for the scenario of storing resources. It would be interesting to extend the presented model to include dimensions without finite value maximising desirability.

In conclusion, our modelling framework maps learning of incentive salience onto the basal ganglia circuitry, a circuitry proven to play an important role in action selection. We used key concepts from both lines of theoretical work to develop a framework that is biologically relevant and describes action selection and learning in a state-dependent manner.

Methods

Simulating state-dependent dopaminergic responses

Here we give details on the simulations of the state dependent dopaminergic response observed in the study by Cone et al. [26] using the classical reward prediction error and the state-dependent reward prediction error. Each simulation consisted of a training phase, in which the expected reinforcements are learned, and a testing phase. Each training phase consisted of 50 trials and one testing trial. The simulations were repeated 5 times (similar to the number of animals used in the study by Cone et al.).

In simulations shown in Fig 3, the prediction error at the time of reinforcement was taken as:

$$\delta = mr - mV \tag{25}$$

where

$$m = \begin{cases} 1, & \text{if } TD, \\ 0.2, & \text{if } balanced, \\ 2, & \text{if } depleted. \end{cases}$$

Even though temporal difference learning is not dependent on the motivational state and therefore does not include the parameter m, we were able to use Eq (25) with $m = 1$ to simulate it. Payoff to the model was $r = 0.5$. During the training phases the expected value of the reinforcement was updated using: $\Delta V_t = \alpha\delta$, where $\alpha = 0.1$. During testing trials, the prediction error at the time of unconditioned stimulus was computed from Eq (25), while at the time of conditioned stimulus was taken as mV.

To illustrate that we can also simulate the state-dependent dopaminergic responses with models grounded in basal ganglia architecture, we also simulated the data by Cone et al. [26] with the gradient and payoff-cost models (Fig 7). All the parameters were kept the same as described above, but the state-dependent prediction error used at the time of reinforcement was computed from Eq (19). For the gradient model we used the plasticity rules described in Eqs (20) and (21). For the payoff-cost model we used the plasticity rules described in Eqs (22) and (23). The parameters used in the simulations were $\alpha = 0.1$, $\epsilon = 0.8$ and $\lambda = 0.01$.

Simulating learning of Go and Nogo weights for different reinforcements and levels of motivation

Simulations shown in Fig 5 were run to compare the ability of models to learn the required values of weights as we differed the level of motivation and reinforcement magnitude. Simulations were performed using the state-dependent prediction error (Eq (19)). The gradient model used the plasticity rules described in Eqs (20) and (21). The payoff-cost model used the plasticity rules described in Eqs (22) and (23). We considered 4 different levels of motivation: variable, low, baseline and high. In the variable motivational condition the model randomly experienced either a low, baseline or high level of motivation, which changed on each trial. Motivation for the other conditions was fixed, where low was $m_{low} = 0$, baseline was $m_{baseline} = 1$ and high was $m_{high} = 2$. For each condition there were four possible reinforcement magnitudes $r \in [0.2, 1, 2, 3]$ which were all simulated independently. Each condition was simulated independently and each simulation consisted of 150 trials and was repeated 100 times. All synaptic weights were initialised at 0.1. The parameters used in the simulations were $\alpha = 0.1$, $\epsilon = 0.8$ and $\lambda = 0.01$. These parameters allow the model to converge to positive and negative consequences at baseline motivational state [42].

Inductive bias allows for faster learning

Simulations in Fig 6 show that the gradient model learns the estimate of the utility faster than standard Q-learning as it relies on prior assumption about the form of the utility. The gradient model was simulated using the state-dependent prediction error (Eq (19)) and the plasticity rules described in Eqs (20) and (21). In our implementation of Q-learning, the range of values the motivation can take was divided into a number of bins. We considered equal sized bins in the range between 0 and 2. To evaluate the influence of discretisation we compared two bin sizes, namely 4 and 10 bins. The model estimated the utility for each bin Q_m. The Q-value for the current motivation bin was updated by: $\Delta Q_m = \alpha(U - Q_m)$. In the simulations on each trial the model received a reinforcement of $r = 1$ and its utility was computed using Eq (4), which relied on the current motivation. The learning rate parameter was also varied across simulations and took values $\alpha \in [0.1, 0.3, 0.5]$.

Influence of physiological state on action selection

In simulations shown in Fig 8, the relevant and irrelevant conditioned stimuli were learned independently in a training phase. For the relevant cue (CS+, sodium), the motivation was set to $m = 0.2$, and for the irrelevant cue (CS-, fructose) the motivation was set to $m = 0.1$. The motivation is encoded in the level of the dopaminergic activation signal following Eq (18). The training phase consisted of 50 trials and was repeated 5 times. The models used the state-dependent prediction error described in Eq (19), the plasticity rules described in Eqs (20) and (21) for the gradient model and the plasticity rules described in Eqs (22) and (23) for the payoff-cost model. The thalamic activity was computed using Eq (14), and additional Gaussian noise ($\mu = 0$ and $\sigma = 0.1$) was added to allow exploration. Actions were made when the thalamic activity was positive, otherwise no action was made. The model received a reinforcement of $r = 0.5$ for each action made and the utility was computed using Eq (13). Once the training phase was completed, the Go and Nogo values were fixed to the learned values and not allowed to be (re-)learned during the testing phase. During the testing phase, the motivation for the depleted state of salt was set to $m = 2$, for the balanced state to $m = 0.2$ and for fructose to $m = 0.1$ (independent of salt deprivation). Using the different levels of dopaminergic activation signal Eq (18) for the different conditions, we computed the thalamic activity and added a

Gaussian noise with $\mu = 0$ and $\sigma = 0.1$. Whenever the activity of an action was positive the action was taken and added to the total number of actions made as depicted in Fig 8.

State-dependent valuation

The simulation shown in Fig 9 was divided into a training and a testing phase. During the training phase, the models learned the values for the option associated with being sated or hungry. These values were learned independently, because only one option was available at the time (forced choice trials). Motivation when sated was set to $m = 0.2$, and when hungry was $m = 2$. The models used the state-dependent prediction error described in (Eq (19)), the plasticity rules described in Eqs (20) and (21) for the gradient model and the plasticity rules described in Eqs (22) and (23) for the payoff-cost model. Each model learned two Go weights and two Nogo weights. The thalamic activity was computed using Eq (14), and additional Gaussian noise ($\mu = 0$ and $\sigma = 0.1$) was added to allow exploration. The number of trials during training phase for each simulated animals was chosen randomly from a normal distribution with a mean 65 and standard deviation 5.5 trials, giving similar number of trials to those in the paper by Aw et al. The animals were tested in either the sated condition or the hungry condition. Motivation for the sated condition, $m = 0.2$, and for the hunger condition, $m = 2$, was used to calculate the corresponding dopamine level using Eq (18). Testing phase consisted of 24 trials during which the individual Go and Nogo weights of option associated with hunger and the sated condition were used to compute a thalamic activity. The training condition that generated the highest thalamic activity was chosen. The proportion of actions for the arm associated with the hungry option were calculated. The simulations (including training and testing) were repeated 11 times (corresponding to the number of animals in the study of Aw et al.), and the average proportions are depicted in Fig 9. The parameters used in the simulations were $\alpha = 0.1$, $\epsilon = 0.8$ and $\lambda = 0.01$.

Fig 10 was generated using the same method and parameters as described above for the training phase, except the number of trials and repetitions was increased so that the simulation was run for 1000 trials and was repeated 100 times.

Acknowledgments

We thank Moritz Möller for his feedback on the manuscript.

Author Contributions

Conceptualization: Rafal Bogacz.

Formal analysis: Maaike M. H. van Swieten.

Investigation: Maaike M. H. van Swieten.

Supervision: Rafal Bogacz.

Writing – original draft: Maaike M. H. van Swieten.

Writing – review & editing: Maaike M. H. van Swieten, Rafal Bogacz.

References

1. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275 (5306):1593–1599. https://doi.org/10.1126/science.275.5306.1593 PMID: 9054347
2. Schultz W. Predictive reward signal of dopamine neurons. Journal of Neurophysiology. 1998; 80(1):1–27. https://doi.org/10.1152/jn.1998.80.1.1 PMID: 9658025

3. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004; 306(5703):1940–1943. https://doi.org/10.1126/science.1102941 PMID: 15528409

4. Wise RA. Dopamine, learning and motivation. Nature Reviews Neuroscience. 2004; 5(6):483–494. https://doi.org/10.1038/nrn1406 PMID: 15152198

5. Frank MJ. Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. Journal of Cognitive Neuroscience. 2005; 17(1):51–72. https://doi.org/10.1162/0898929052880093 PMID: 15701239

6. Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 2005; 47(1):129–141. https://doi.org/10.1016/j.neuron.2005.05.020 PMID: 15996553

7. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 2006; 442(7106):1042–1045. https://doi.org/10.1038/nature05051 PMID: 16929307

8. Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. A causal link between prediction errors, dopamine neurons and learning. Nature Neuroscience. 2013; 16(7):966–973. https://doi.org/10.1038/nn.3413 PMID: 23708143

9. Berridge KC, Robinson TE. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Research Reviews. 1998; 28(3):309–369. https://doi.org/10.1016/S0165-0173(98)00019-8 PMID: 9858756

10. Niv Y, Daw ND, Joel D, Dayan P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology. 2007; 191(3):507–520. https://doi.org/10.1007/s00213-006-0502-4 PMID: 17031711

11. Robbins TW, Everitt BJ. A role for mesencephalic dopamine in activation: commentary on Berridge (2006). Psychopharmacology. 2007; 191(3):433–437. https://doi.org/10.1007/s00213-006-0528-7 PMID: 16977476

12. Berridge KC. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology. 2007; 191(3):391–431. https://doi.org/10.1007/s00213-006-0578-x PMID: 17072591

13. Shen W, Flajolet M, Greengard P, Surmeier DJ. Dichotomous dopaminergic control of striatal synaptic plasticity. Science (New York, NY). 2008; 321(5890):848–851. https://doi.org/10.1126/science.1160575

14. Tobler PN, Fiorillo CD, Schultz W. Adaptive coding of reward value by dopamine neurons. Science. 2005; 307(5715):1642–1645. https://doi.org/10.1126/science.1105370 PMID: 15761155

15. Fiorillo CD, Tobler PN, Schultz W. Discrete coding of reward probability and uncertainty by dopamine neurons. Science. 2003; 299(5614):1898–1902. https://doi.org/10.1126/science.1077349 PMID: 12649484

16. Day JJ, Roitman MF, Wightman RM, Carelli RM. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nature Neuroscience. 2007; 10(8):1020–1028. https://doi.org/10.1038/nn1923 PMID: 17603481

17. Gan JO, Walton ME, Phillips PEM. Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. Nature neuroscience. 2010; 13(1):25–27. https://doi.org/10.1038/nn.2460 PMID: 19904261

18. Thurley K, Senn W, Lüscher HR. Dopamine increases the gain of the input-output response of rat prefrontal pyramidal neurons. Journal of Neurophysiology. 2008; 99(6):2985–2997. https://doi.org/10.1152/jn.01098.2007 PMID: 18400958

19. Hernández-López S, Tkatch T, Perez-Garci E, Galarraga E, Bargas J, Hamm H, et al. D2 dopamine receptors in striatal medium spiny neurons reduce L-Type Ca2+ currents and excitability via a novel PLCβ1-IP3-Calcineurin-signaling cascade. Journal of Neuroscience. 2000; 20(24):8987–8995. https://doi.org/10.1523/JNEUROSCI.20-24-08987.2000 PMID: 11124974

20. Moyer JT, Wolf JA, Finkel LH. Effects of Dopaminergic Modulation on the Integrative Properties of the Ventral Striatal Medium Spiny Neuron. Journal of Neurophysiology. 2007; 98(6):3731–3748. https://doi.org/10.1152/jn.00335.2007 PMID: 17913980

21. da Silva JA, Tecuapetla F, Paixão V, Costa RM. Dopamine neuron activity before action initiation gates and invigorates future movements. Nature. 2018; 554(7691):244–248. https://doi.org/10.1038/nature25457 PMID: 29420469

22. Mohebi A, Pettibone JR, Hamid AA, Wong JMT, Vinson LT, Patriarchi T, et al. Dissociable dopamine dynamics for learning and motivation. Nature. 2019; 570(7759):65–70. https://doi.org/10.1038/s41586-019-1235-y PMID: 31118513

23. Berke JD. What does dopamine mean? Nature Neuroscience. 2018; 21:787–793. https://doi.org/10.1038/s41593-018-0152-y PMID: 29760524

24. Aitken TJ, Greenfield VY, Wassum KM. Nucleus accumbens core dopamine signaling tracks the need-based motivational value of food-paired cues. Journal of neurochemistry. 2016; 136(5):1026–1036. https://doi.org/10.1111/jnc.13494 PMID: 26715366

25. Papageorgiou GK, Baudonnat M, Cucca F, Walton ME. Mesolimbic dopamine encodes prediction errors in a state-dependent manner. Cell Reports. 2016; 15(2):221–228. https://doi.org/10.1016/j.celrep.2016.03.031 PMID: 27050518

26. Cone JJ, Fortin SM, McHenry JA, Stuber GD, McCutcheon JE, Roitman MF. Physiological state gates acquisition and expression of mesolimbic reward prediction signals. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113(7):1943–1948. https://doi.org/10.1073/pnas.1519643113 PMID: 26831116

27. Berridge KC, Schulkin J. Palatability shift of a salt-associated incentive during sodium depletion. The Quarterly journal of experimental psychology B, Comparative and physiological psychology. 1989; 41 (2):121–138. PMID: 2748936

28. Aw JM, Holbrook RI, Burt de Perera T, Kacelnik A. State-dependent valuation learning in fish: Banded tetras prefer stimuli associated with greater past deprivation. Behavioural Processes. 2009; 81(2):333–336. https://doi.org/10.1016/j.beproc.2008.09.002 PMID: 18834933

29. Keramati M, Gutkin B. Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife. 2014; 3:e04811. https://doi.org/10.7554/eLife.04811

30. Stigler GIJ. The development of utility theory; 1950. 4. Available from: https://www.jstor.org/stable/pdf/1828885.pdf?refreqid=excelsior{%}3Aa6bbd6a0753192d5f7948035ab0a5268.

31. Stauffer WR, Lak A, Schultz W. Dopamine reward prediction error responses reflect marginal utility. Current biology: CB. 2014; 24(21):2491–500. https://doi.org/10.1016/j.cub.2014.08.064 PMID: 25283778

32. Dickinson A, Balleine B. The role of learning in the operation of motivational systems. In: Stevens' Handbook of Experimental Psychology. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2002. p. 497–534. Available from: http://doi.wiley.com/10.1002/0471214426.pas0312.

33. Levy DJ, Thavikulwat AC, Glimcher PW. State dependent valuation: The effect of deprivation on risk preferences. PLoS ONE. 2013; 8(1):e53978. https://doi.org/10.1371/journal.pone.0053978 PMID: 23358126

34. Kacelnik A, Marsh B. Cost can increase preference in starlings. Animal Behaviour. 2002; 63(2):245–250. https://doi.org/10.1006/anbe.2001.1900

35. Pompilio L, Kacelnik A. State-dependent learning and suboptimal choice: when starlings prefer long over short delays to food. Animal Behaviour. 2005; 70(3):571–578. https://doi.org/10.1016/j.anbehav.2004.12.009

36. Pompilio L, Kacelnik A, Behmer ST. State-dependent learned valuation drives choice in an invertebrate. Science. 2006; 311(5767):1613–1615. https://doi.org/10.1126/science.1123924 PMID: 16543461

37. McClure SM, Daw ND, Read Montague P. A computational substrate for incentive salience. Opinion TRENDS in Neurosciences. 2003; 26(8):423–428. https://doi.org/10.1016/S0166-2236(03)00177-2

38. Zhang J, Berridge KC, Tindell AJ, Smith KS, Aldridge JW. A neural computational model of incentive salience. PLoS Computational Biology. 2009; 5(7):e1000437. https://doi.org/10.1371/journal.pcbi.1000437 PMID: 19609350

39. Kravitz AV, Freeze BS, Parker PRL, Kay K, Thwin MT, Deisseroth K, et al. Regulation of parkinsonian motor behaviors by optogenetic control of basal ganglia circuitry. Nature. 2010; 466(7306):622. https://doi.org/10.1038/nature09159 PMID: 20613723

40. Smith Y, Bevan MD, Shink E, Bolam JP. Microcircuitry of the direct and indirect pathways of the basal ganglia. Neuroscience. 1998; 86(2):353–387. PMID: 9881853

41. Surmeier DJ, Ding J, Day M, Wang Z, Shen W. D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends in Neurosciences. 2007; 30(5):228–235. https://doi.org/10.1016/j.tins.2007.03.008 PMID: 17408758

42. Möller M, Bogacz R. Learning the payoffs and costs of actions. PLOS Computational Biology. 2019; 15 (2):e1006285. https://doi.org/10.1371/journal.pcbi.1006285 PMID: 30818357

43. Lee HJ, Weitz AJ, Bernal-Casas D, Duffy BA, Choy MK, Kravitz AV, et al. Activation of Direct and Indirect Pathway Medium Spiny Neurons Drives Distinct Brain-wide Responses. Neuron. 2016; 91(2):412–424. https://doi.org/10.1016/j.neuron.2016.06.010 PMID: 27373834

44. Dreyer JK, Herrik KF, Berg RW, Hounsgaard JD. Influence of phasic and tonic dopamine release on receptor activation. Journal of Neuroscience. 2010; 30(42):14273–14283. https://doi.org/10.1523/JNEUROSCI.1894-10.2010 PMID: 20962248

45. Marcott PF, Mamaligas AA, Ford CP. Phasic dopamine release drives rapid activation of striatal D2-receptors. Neuron. 2014; 84(1):164–176. https://doi.org/10.1016/j.neuron.2014.08.058 PMID: 25242218

46. Yapo C, Nair AG, Clement L, Castro LR, Hellgren Kotaleski J, Vincent P. Detection of phasic dopamine by D1 and D2 striatal medium spiny neurons. Journal of Physiology. 2017; 595(24):7451–7475. https://doi.org/10.1113/JP274475 PMID: 28782235

47. Hunger L, Kumar A, Schmidt R. Abundance compensates kinetics: Similar effect of dopamine signals on D1 and D2 receptor populations. The Journal of Neuroscience. 2020; p. 1951–19.

48. Gurney K, Prescott TJ, Redgrave P. A computational model of action selection in the basal ganglia. I. A new functional anatomy. Biological Cybernetics. 2001; 84(6):401–410. https://doi.org/10.1007/PL00007985 PMID: 11417052

49. Mikhael JG, Bogacz R. Learning reward uncertainty in the basal ganglia. PLOS Computational Biology. 2016; 12(9):e1005062. https://doi.org/10.1371/journal.pcbi.1005062 PMID: 27589489

50. Mitchell TM. Machine learning. 1st ed. New York, NY, USA: McGraw-Hill Inc.; 1997.

51. Krieckhaus EE, Wolf G. Acquisition of sodium by rats: Interaction of innate mechanisms and latent learning. Journal of Comparative and Physiological Psychology. 1968; 65(2):197–201. PMID: 5668303

52. Marsh B, Schuck-Paim C, Kacelnik A. Energetic state during learning affects foraging choices in starlings. Behavioral Ecology. 2004; 15(3):396–399. https://doi.org/10.1093/beheco/arh034

53. Berridge KC. From prediction error to incentive salience: mesolimbic computation of reward motivation. The European journal of neuroscience. 2012; 35(7):1124–1143. https://doi.org/10.1111/j.1460-9568.2012.07990.x PMID: 22487042

54. Shin JH, Kim D, Jung MW. Differential coding of reward and movement information in the dorsomedial striatal direct and indirect pathways. Nature Communications. 2018; 9(1):1–14. https://doi.org/10.1038/s41467-017-02817-1

55. Weismüller B, Ghio M, Logmin K, Hartmann C, Schnitzler A, Pollok B, et al. Effects of feedback delay on learning from positive and negative feedback in patients with Parkinson's disease off medication. Neuropsychologia. 2018; 117:46–54. https://doi.org/10.1016/j.neuropsychologia.2018.05.010 PMID: 29758227

56. Siep N, Roefs A, Roebroeck A, Havermans R, Bonte ML, Jansen A. Hunger is the best spice: An fMRI study of the effects of attention, hunger and calorie content on food reward processing in the amygdala and orbitofrontal cortex. Behavioural Brain Research. 2009; 198(1):149–158. https://doi.org/10.1016/j.bbr.2008.10.035 PMID: 19028527

57. Abizaid A, Liu ZW, Andrews ZB, Shanabrough M, Borok E, Elsworth JD, et al. Ghrelin modulates the activity and synaptic input organization of midbrain dopamine neurons while promoting appetite. The Journal of clinical investigation. 2006; 116(12):3229–3239. https://doi.org/10.1172/JCI29867 PMID: 17060947

58. van der Plasse G, van Zessen R, Luijendijk MCM, Erkan H, Stuber GD, Ramakers GMJ, et al. Modulation of cue-induced firing of ventral tegmental area dopamine neurons by leptin and ghrelin. International Journal of Obesity. 2015; 39(12):1742–1749. https://doi.org/10.1038/ijo.2015.131 PMID: 26183405

59. Iodice P, Ferrante C, Brunetti L, Cabib S, Protasi F, Walton ME, et al. Fatigue modulates dopamine availability and promotes flexible choice reversals during decision making. Scientific Reports. 2017; 7(1):535. https://doi.org/10.1038/s41598-017-00561-6 PMID: 28373651

60. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science. 2004; 304(5669):452–454. https://doi.org/10.1126/science.1094285 PMID: 15087550

61. Symmonds M, Emmanuel JJ, Drew ME, Batterham RL, Dolan RJ. Metabolic state alters economic decision making under risk in humans. PloS one. 2010; 5(6):e11090. https://doi.org/10.1371/journal.pone.0011090 PMID: 20585383

62. Wang GJ, Volkow ND, Logan J, Pappas NR, Wong CT, Zhu W, et al. Brain dopamine and obesity. Lancet (London, England). 2001; 357(9253):354–7. https://doi.org/10.1016/S0140-6736(00)03643-6

63. Stice E, Spoor S, Bohon C, Small DM. Relation between obesity and blunted striatal response to food is moderated by TaqIA A1 allele. Science (New York, NY). 2008; 322(5900):449–52. https://doi.org/10.1126/science.1161550

64. Wang GJ, Volkow ND, Thanos PK, Fowler JS. Imaging of brain dopamine pathways: implications for understanding obesity. Journal of addiction medicine. 2009; 3(1):8–18. https://doi.org/10.1097/ADM.0b013e31819a86f7 PMID: 21603099

65. Friston K. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience. 2010; 11(2):127–138. https://doi.org/10.1038/nrn2787 PMID: 20068583

66. Stephan KE, Manjaly ZM, Mathys CD, Weber LAE, Paliwal S, Gard T, et al. Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. Frontiers in Human Neuroscience. 2016; 10:550. https://doi.org/10.3389/fnhum.2016.00550 PMID: 27895566

67. Collins AGE, Frank MJ. Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychological Review. 2014; 121(3):337–366. PMID: 25090423

68. Dunovan K, Vich C, Clapp M, Verstynen T, Rubin J. Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making. PLOS Computational Biology. 2019; 15(5): e1006998. https://doi.org/10.1371/journal.pcbi.1006998 PMID: 31060045