Dorsal anterior cingulate-midbrain ensemble as a reinforcement meta-learner

Massimo Silvetti[1*], Eliana Vassena[1,2], Tom Verguts[1]

1 Ghent University, Department of Experimental Psychology, Ghent, Belgium
2 Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

* Corresponding author:
massimo.silvetti@ugent.be

key words: ACC, VTA, LC, reinforcement learning, meta-learning, Bayesian learning, Kalman filter, effort control, higher-order conditioning.

Abstract

The dorsal anterior cingulate cortex (dACC) is central in higher-order cognition and in the pathogenesis of several mental disorders. Reinforcement Learning (RL), Bayesian decision-making, and cognitive control are currently the three main theoretical frameworks attempting to capture the elusive computational nature of this brain area. Although theoretical effort to explain the dACC functions is intense, no single theoretical framework managed so far to account for the myriad of relevant experimental data. Here we propose that dACC plays, in concert with midbrain catecholamine nuclei, the role of a reinforcement meta-learner. This cortical-subcortical system not only can learn and make decisions based on RL principles, but it can also learn to control the learning process itself, for both its own circuits and for other brain areas. We show that a neural model implementing this theory, the Reinforcement Meta-Learner (RML), can account for an unprecedented number of experimental findings among which effort exertion, higher-order conditioning and working memory. The RML performs meta-learning by means of approximate Bayesian inference and it respects several neuro-functional and neuro-anatomical constraints, providing a perspective that assimilates the other theoretical proposals in a single computational framework.

Introduction

Adapting behavior to uncertain and changing environments is the foundation of intelligence. Important theoretical progress was made by considering this behavioural adaptation as a problem of decision-making (Frank et al., 2004; Rushworth and Behrens, 2008). At anatomical-functional level, the dorsal anterior cingulate cortex (dACC) was proposed as a multifunctional hub with a pivotal role in decision-making (Rushworth and Behrens, 2008). In recent years, Reinforcement Learning (RL) neural models (Silvetti et al., 2014) showed that many signals recorded in the dACC (e.g., error, error likelihood, response conflict) can be explained in terms of reward expectation and prediction error (PE) (i.e. the difference between expectation and outcome) for the purpose of optimal decision-making. This framework proposed that the dACC is a multi-domain estimator of stimulus and action values for the dACC, aimed at maximizing long-term reward.

Nonetheless, other recent studies showed that the function of dACC extends beyond the role of expectation-outcome comparator, revealing its capability of adaptive control over internal parameters determining behaviour. For example, from a cognitive control perspective it has been proposed that dACC controls effort exertion (Croxson et al., 2009; Shenhav et al., 2013; Vassena et al., 2014; Verguts et al., 2015). Further, from a Bayesian perspective the dACC controls learning rate to optimize behavioural adaptation (Behrens et al., 2007; Kolling et al., 2016). Despite several recent theoretical proposals on the role of dACC in RL, cognitive control and Bayesian adaptation (Ebitz and Hayden, 2016), no theoretical convergence was reached yet and no concrete computational model has been developed to reconcile (or allow competition between) the different theoretical positions.

Here we propose a new perspective on dACC function, which is that the dACC performs meta-learning operations. From this perspective, the dACC learns to control the parameters that drive learning itself and decision-making, resulting in a more flexible system, better capable of adapting to the environment (Doya 2002; kahmassi 2011). In addition to dACC, midbrain catecholamine nuclei (the ventral tegmental area (VTA) and the locus coeruleus (LC)) probably play a crucial role in meta-learning as well (Doya, 2002). Dopamine (DA) and norepinephrine (Ne) are involved in control over both physical and cognitive effort exertion (Salamone et al.,

1994; Walton et al., 2009; Varazzani et al., 2015). Further, Ne levels have been linked with promoting knowledge update in case of environmental changes, balancing the trade-off between knowledge stability and plasticity (Nassar et al., 2012; Silvetti et al., 2013a).

We propose that the heterogeneity of dACC signals can be explained by considering it as a subsystem of a broader cortical-subcortical circuit involved in learning and meta-learning, including the dACC itself and the midbrain catecholamine nuclei. We implemented this hypothesis in a novel RL neural model, the Reinforcement Meta-Learner (RML). The RML goal is to maximize reward not only by making decisions toward the external environment, but also by self-modulating catecholamines release, thus to modify (i.e., meta-learn) its own parameters. We show how this model has sufficient flexibility for RL, cognitive control and (approximate) Bayesian adaptation to emerge in complete autonomy (free from the intervention of the "homunculus"), in a (systems-level) neurophysiologically plausible way. The RML can also work as a central "server" of cognitive control, providing control signals to other brain areas to optimize performance. The RML accounts for an unprecedented variety of data, including volatility estimation, effort processing, and working memory (WM), higher-order classical and instrumental conditioning.

In the next sections, we briefly describe the model structure. Then, we present the results on both neural and behavioural dynamics of the RML in nine key experimental paradigms, selected based on their widely accepted effectiveness in testing RL, cognitive control and Bayesian adaptation phenomena.

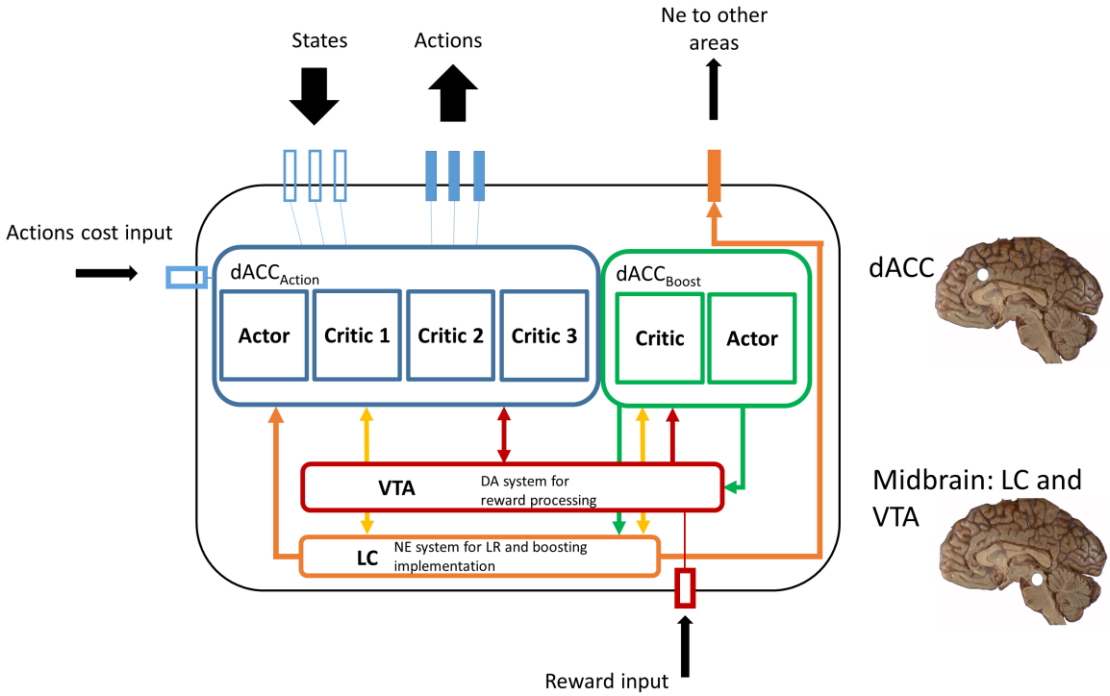

Figure 1. Model overview. The RML has nine channels of information exchange (black arrows) with the environment (input = empty bars; output = filled bars). The input channels consist of one channel encoding action costs, three encoding environmental states and one encoding primary rewards. The output consists of three channels coding each for one action, plus one channel conveying LC signals to other brain areas. The entire model is composed of four reciprocally connected modules (each a different color). The upper modules (blue and green) simulate the dACC, while the lower modules (red and orange) simulate the midbrain catecholamine nuclei (VTA and LC). dACC$_{Action}$ selects actions directed toward the environment and learns higher-order conditioning (each Critic is dedicated to one order of conditioning). dACC$_{Boost}$ selects actions directed toward the catecholamine nuclei, modulating their output. Both dACC modules are Actor-Critic systems. The VTA module provides DA training signals to both dACC modules, while the LC controls learning rate (yellow bidirectional arrow) in both dACC modules, and costs estimation in the dACC$_{Action}$ module (orange arrow), influencing their decisions. Finally, the same signal controlling costs estimation in the dACC$_{Action}$ is directed toward other brain areas for neuro-modulation.

Model description

In this section, we describe the general RML principles; for a detailed description, the reader can consult the Supplementary Methods. The model approximates optimal decision-making (to maximize long-term reward) based on action-outcome comparisons (by the dACC; blue and green modules in Figure 1) augmented by meta-learning (LC and VTA; orange and red modules in Figure 1). It does this via value estimation of states (*s*) and actions (*a*), which are subsequently used for decision making (Sutton and Barto, 1998). Communication with the external environment is based on 9 channels. There are six channels representing environmental states and RML actions (3 states and 3 actions). The first two actions are aimed at changing the environmental state (e.g. turning right or left), while the 3rd action means "Stay", i.e. refusing to engage in the task. This action is never rewarded and has no associated costs. Neural units are modeled by stochastic leaky integrators simulating neural activity with time resolution of 10 ms (the time length assigned to each network cycle). The model embeds several copies of our previous RL model RVPM (Silvetti et al., 2011). Each RVPM module functions as one Critic (in both dACC modules), which are involved in both action selection and mid-brain modulation. Our model is scalable by design, i.e. there is no theoretical limit to the number of critic submodules and state/action channels, while the number of parameters does not change as a function of model size.

dACC$_{Action}$

The dACC$_{Action}$ module consists of a network of Critics and one Actor. Each Critic is a performance evaluator and computes reward expectation and PE for either primary or non-primary rewards (higher order conditioning), learning to associate stimuli and actions to environmental outcomes. We simulated a hierarchy of abstraction levels for higher-order conditioning, up to 3rd order. Each abstraction level is learned by a Critic. The Critic at the lowest level learns from primary rewards. The higher-order Critics learn by signals from lower level Critics, signals that are first processed by the VTA (Figure 1; see VTA module description below). The Actor selects motor actions (based on Critics expectation) to maximize long-term reward.

The central equation in this module governs (any) Critic state/action value update:

$$\Delta v_t(s,a) = \lambda\left(B_t - T v_{t-1}(s,a)\right) \quad (1)$$

where $v(s,a)$ indicates the value (outcome prediction) of a specific action a given a state s. In a continuous time representation, T is a timing function modeling the outcome onset expectation (Silvetti et al., 2011). T can be either set or autonomously learned (Silvetti et al., 2013b). A dedicated signal for time representation allows to go beyond some important limitations of temporal difference learning algorithms, as we describe in the General Discussion and in the Supplementary Methods. Equation 1 ensures that v comes to resemble the outcome (B) as veridically as possible. It entails that the update of v at time step t is based on the difference between prediction (v) and outcome (B), which defines the concept of PE. The latter is weighted by learning rate λ, making the update more (high lambda) or less (low lambda) dependent on recent events. In the most general case, outcome B is defined by a reward function f:

$$B_t = f\left(r_t, b\right) \quad (2)$$

where the output f of the VTA module (see Equation 7) is a function of r (the external feedback signal) with parameter b, regulating its gain. The meaning of r can be either a primary reward or a conditioned cue (in case of higher-order conditioning; Equation 7).

Action a is selected by the Actor subsystem, which implements softmax action selection (with temperature τ) based on state/action values discounted by action costs C:

$$p(a \mid s) = \mathrm{softmax}\left(v(s,a) - C(a,\zeta),\tau\right) \quad (3)$$

Regardless of the algebraic form of f and C, these functions contain parameters (b and ζ) that (together with learning rate λ) are typically fixed and optimized before task engagement. However, this approach suffers from limited flexibility. The RML solves this problem by autonomously modulating parameters λ, ζ, and b. First, modulation

of *learning rate* (λ) can make sure knowledge (stored in *v*) is updated only when there are relevant environmental changes, protecting it from non-informative random fluctuations. This addresses the classical stability-plasticity trade-off (Grossberg, 1980). Second, modulation of *costs* estimation (by control over ζ) allows optimization when the benefit-cost balance changes over time and/or it is particularly challenging to estimate (e.g. when it is necessary to pay a sure high cost to get an uncertain higher reward). Third, dynamic modulation of *reward* signals (by means of control over *b*) is the foundation for emancipating learning from primary rewards (see Equation 7 and Equation s5), allowing the RML to learn complex tasks without the immediate availability of primary rewards (higher order conditioning).

dACC$_{boost}$

The dACC$_{Boost}$ module consists of one Critic and one Actor. This module controls the parameters for cost and reward signals in equations 2-3 (dACC$_{Action}$), via modulation of VTA and LC activity (boosting catecholamines). In other words, whereas the dACC$_{Action}$ decides on actions toward the external environment, the dACC$_{Boost}$ decides on actions toward the internal environment: It modulates midbrain nuclei (VTA and LC), given a specific environmental state. This is implemented by selecting the modulatory signal *b* (*boost signal*), by RL-based decision-making. In our model, *b* is a discrete signal that can assume ten different values (integers 1-10), each corresponding to one action selectable by the dACC$_{Boost}$. The Critic submodule inside the dACC$_{Boost}$ estimates the boost values via the equation:

$$\Delta v_t(s,b) = \lambda\left(r - \chi(b) - Tv_{t-1}(s,b)\right) \tag{4}$$

Equation 4 represents the value update of boosting level *b* in the environmental state *s*. The Actor submodule selects boosting actions to maximize long-term reward. Function $\chi(b)$ represents the cost of boosting (Kool et al., 2010; Kool and Botvinick, 2013; Shenhav et al., 2013). Referring to Equation 2, the dACC$_{Boost}$ modulates the reward signal by changing the parameter *b* in function $f(r,b)$, where function *f* is coded in VTA. Furthermore, dACC$_{Boost}$ also modulates the cost signal by changing parameter ζ (via LC module, Equation 5) in the function

representing action cost $C(a,\zeta)$ (Equation 3; represented in the Actor within the dACC$_{Action}$).

LC

The LC module plays a double role. First it controls cost via parameter ζ, as a function of boosting value b selected by the dACC$_{Boost}$ module:

$$\zeta = LC(b) \tag{5}$$

Parameter ζ is modulated by boosting b, via the monotonically increasing LC function. The cost C decreases when ζ increases, so action costs C decrease when b increases. The Ne output (ζ) is directed also toward external brain areas as a performance modulation signal (Figure 2; Simulation 2b).

In addition to cost, the LC module dynamically optimizes learning rate (λ) in the two dACC modules. The approximation of optimal λ solves the trade-off between stability and plasticity, increasing learning speed when the environment changes and lowering it when the environment is simply noisy. In this way, the RML updates its knowledge when needed (plasticity), protecting it from random fluctuations. This function is performed by means of recurrent connections between each Critic-unit and the LC module. The resulting algorithm approximates Kalman filtering (Kalman, 1960; Welch and Bishop, 1995), which is a recursive Bayesian estimator. In its simplest formulation, Kalman filter computes expectations (posteriors) from current estimates (priors) plus PE weighted by an adaptive learning rate (called Kalman gain). If we define process variance as the outcome variance due to volatility of the environment, Kalman filter computes the Kalman gain as the ratio between process variance and total variance (i.e. the sum of process and noise variance). From the Bayesian perspective, the Kalman gain reflects the confidence about the prior, so that high values reflect low confidence on priors and more influence by evidence on posteriors estimation. The RML approximates this ratio based on environmental outcomes, without knowing a priori neither process nor noise variance. In order to do

that, the LC modulates λ as a function of the ratio between the estimated variance of state/action-value ($\hat{\mathrm{Var}}(v)$) over the estimated squared PE ($\hat{U}$):

$$\lambda_t(s,a) = \frac{\hat{\mathrm{Var}}(v)_t}{\left(\hat{U}_t(s,a)\right)^2} \tag{6}$$

In Equation 6, process variance is approximated by $\hat{\mathrm{Var}}(v)$, total variance by squared PE (which is due to both noise and volatility). The LC module computes $\hat{\mathrm{Var}}(v)$, with the minimal assumption that noise-related variability occurs at a faster time scale than volatility-related variability (Equation s15 in Supplementary Methods). Equation 6 is implemented independently for each of the Critic submodules in the two dACC modules, so that each Critic (equations 1,4) interacts with the LC to modulate its own learning rate. The dACC modules and the LC play complementary roles in controlling λ: The dACC modules provide the LC with the time course of expectations and PEs occurring during a task, while the LC integrates them to compute Equation 6.

VTA

The VTA provides training signal $f(r,b)$ to both dACC modules, either for action selection directed toward the environment (by $\mathrm{dACC}_{\mathrm{Action}}$) or for boosting-level selection (by $\mathrm{dACC}_{\mathrm{Boost}}$) directed to the midbrain catecholamine nuclei. The VTA module also learns to link dopamine signals to arbitrary environmental stimuli (non-primary rewards) to allow higher-order conditioning. We hypothesize that this mechanism is based on DA shifting from primary reward onset to conditioned stimulus (s, a, or both) onset (Ljungberg et al., 1992). As in the earlier model (Silvetti et al., 2011), DA shifting from reward onset to cue onset is modeled by a combination of the time derivatives of the neural units in $\mathrm{dACC}_{\mathrm{Action}}$ (Silvetti et al., 2011):

$$f(r,b) \propto b\left(\left[\dot{v}\right]^+ + \left[\dot{\delta}^+\right]^+ - \left[\dot{\delta}^-\right]^+\right) \tag{7}$$

where v indicates the output of the value unit, δ^+ and δ^- indicate respectively positive and negative PE activity and $\left[x\right]^+ \equiv \max(0,x)$. Temporal derivatives were chosen to simulate transient VTA activation (Cohen et al., 2012), and rectifications prevent

negative neural input when derivatives are negative. Equation 7 plays a role similar to a temporal difference (TD) signal (Schultz et al., 1997); for detailed comparison between our algorithm and TD-learning see the Discussion section.

Client/server system

Finally, the RML can optimize performance of other brain areas. It does so via the LC-based control signal, which is the same signal that modulates costs (Equation 5; Figure 2). Indeed, the Actor-Critic function of the $dACC_{Action}$ module is domain-independent (i.e. the state/action channels can come from any brain area outside dACC), and this allows a dialogue with any other area. Moreover, because optimization of any brain area improves behavioural performance, the $dACC_{Boost}$ can modulate (via LC signals defined by Equation 5) any cortical area to improve performance.

Example of RML dynamics

To visualize the RML functions, we show in Figure 3 the state-action transitions during a trial in a higher-order instrumental conditioning paradigm (a simplified version of the task described in Figure 10), where the RML needs to perform two different actions (and transit through two different environment states) before obtaining the final reward. As we will show in more detail in Simulation 3, the transition to a new state, closer to primary reward, plays the role of a non-primary reward that is used by the RML to update the value of the action that determined the transition to that state.

In the following sections, we describe seven different simulations where we show how the RML implements meta-learning, controlling autonomously (in order of simulation description): learning rate (λ), physical effort exertion (modulation of action costs C), modulation of the dorsolateral prefrontal cortex (DLPFC) via LC efferent signals (ζ), and reward signals for higher-order conditioning (b, Equation 7). Across all simulations, the exact same architecture and parameter setting was used.

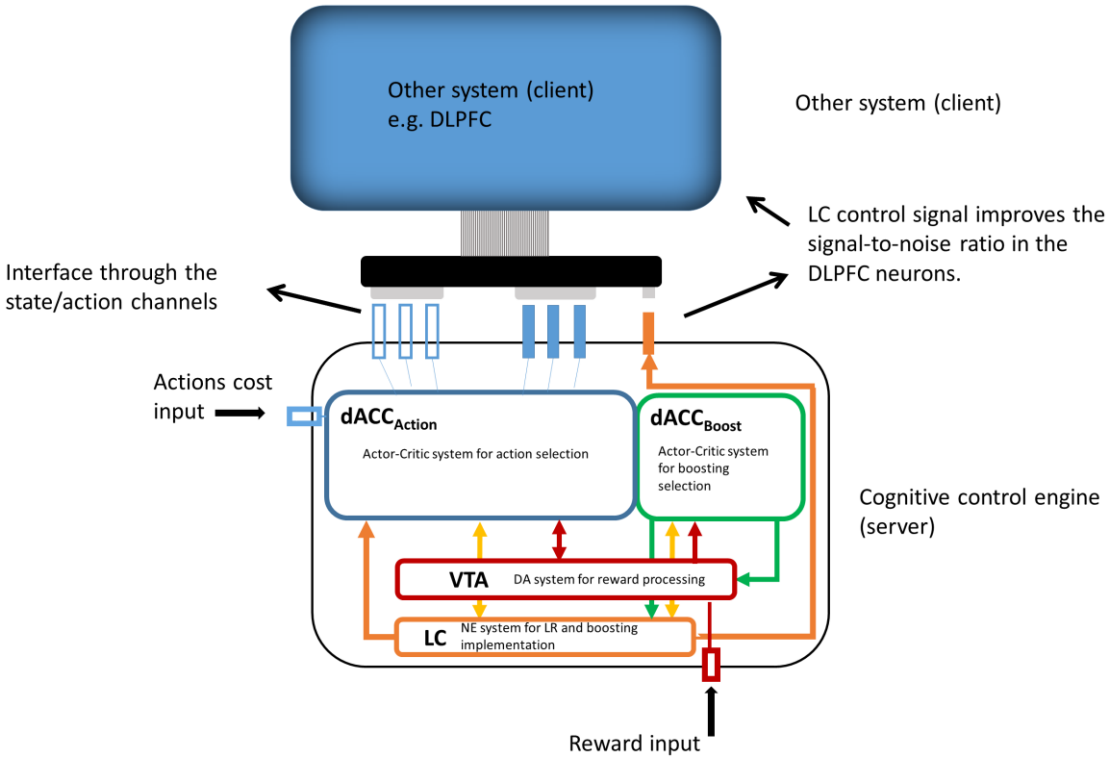

Figure 2. Example of how the RML can provide control signals to other brain areas. The schema represents the connection of one external system (client; e.g. the DLPFC) to the RML (server). The communication between the two systems is ensured by the blue state/action channels. The LC module provides external control signal, to modulate DLPFC activity (e.g. signal-to-noise ratio). Boosting LC activation is decided by the $dACC_{Boost}$ module to maximize long-term reward (see also Simulation 2b).

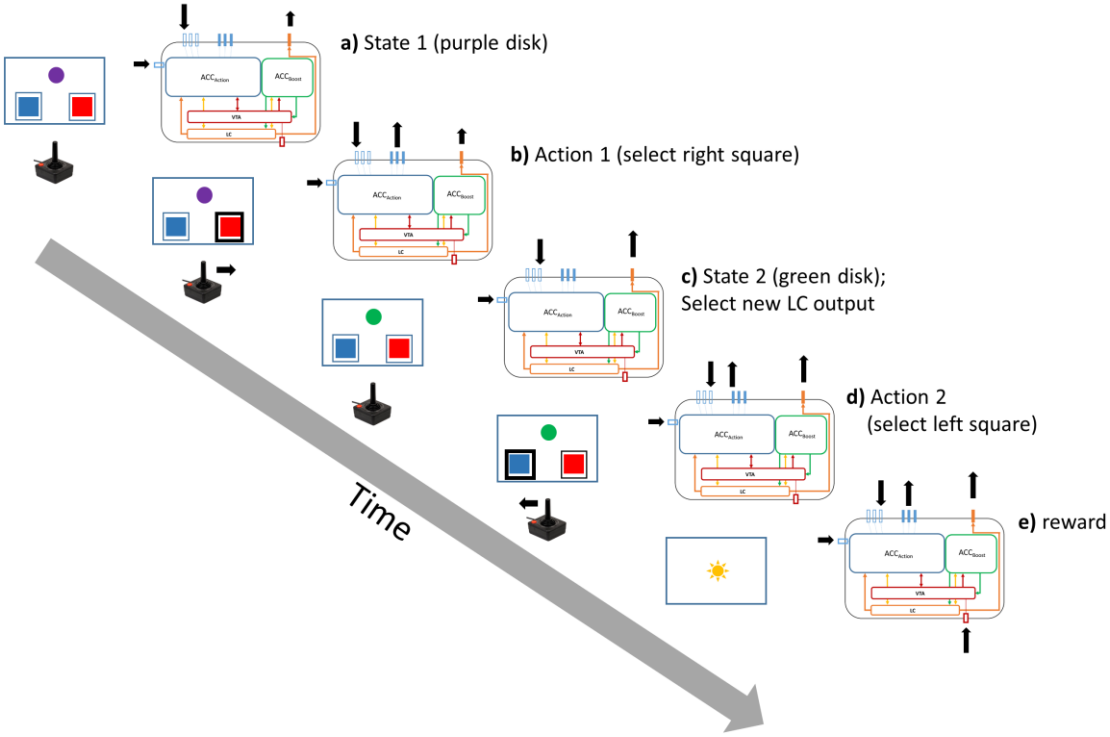

Figure 3. Example of input-output RML dynamics (right sequence) during a trial in a higher-order instrumental conditioning task (left sequence). The task is represented like a 2-armed bandit task, where two options (red and blue squares) can be selected (joystick). The RML needs to select a sequence of two actions before achieving a primary reward (sun). Each action determines an environmental state transition (colored disks) **a)** Trial starts in environmental State 1 (purple disk on selection screen). This state is encoded in the state input layer of the RML (black arrow to the first RML input pin), while the LC output level to external brain areas (same level of internal LC output) has been selected based on prior knowledge (black arrow from LC output). **b)** The RML decides to select the right gate (black arrow from the second RML output pin and black arrow indicating joystick movement). **c)** The environment changes state (green disk on display) as a consequence of the RML action. The new environment state is encoded by the RML (black arrow to the third RML input pin). Based on the new environment state, the RML makes a new decision about the LC output intensity (longer black arrow from LC output pin). As before, the new LC output will influence both the RML itself and external brain areas receiving the LC output. **d)** The RML selects a new action (left gate). **e)** The environment transits to the final state of the trial, reward (sun), which is received by the RML (black arrow to the RML primary reward pin). In this example, action cost estimations (black arrow to the action cost RML input pin) remain constant during the trial.

Results

Simulation 1: learning rate and Bayesian inference

Adaptive control of learning rate is a fundamental aspect of cognition. Humans can solve the tradeoff between stability and plasticity in a Bayesian fashion, by changing the learning rate as a function of environmental changes (Behrens et al., 2007; Yu, 2007), and distinguishing between variability due to noise versus variability due to actual changes of the environment (Yu and Dayan, 2005; Silvetti et al., 2013a). The RML implements learning rate meta-learning by means of recurrent interaction between the dACC and the LC (Equation s15, see also Suppl. Material), allowing it to estimate these quantities too, thus to approximate optimal control.

Furthermore, we will investigate not only whether the model can capture and explain human adaptive control of learning rate at behavioural level, but also a set of experimental findings at neural level, which have not been reconciled yet under one single theoretical framework. In particular, LC activity (and thus Ne release) has been shown to track volatility (probably controlling learning rate); in sharp contrast, dACC activation was more sensitive to global environmental uncertainty, rather than to volatility (Nassar et al., 2012; Silvetti et al., 2013a, 2013b).

Simulation methods

We administered to the RML a 2-armed bandit task in three different stochastic environments (Figure 4a-b, see also Suppl. Material). The three environments were: stationary environment (Stat, where the links between reward probabilities and options were stable over time, either 70 or 30%), stationary with high uncertainty (Stat2, also stable reward probabilities, but all the options led to a reward in 60% of times), and volatile (Vol, where the links between reward probabilities and options randomly changed over time) (see also Table s2). We assigned higher reward magnitudes to choices with lower reward probability, to promote switching between choices and to make the task more challenging (cf. Behrens et al. 2007). Nonetheless, the value of each choice (probability × magnitude) remained higher for higher reward probability (see Supplementary Methods for details), meaning that reward probability was the relevant variable to be tracked. A second experiment, where we manipulated

reward magnitude instead of reward probability (Suppl. Results and Methods), led to very similar results. Here and elsewhere, to mimic standard experimental paradigms as closely as possible, we ran just 12 simulations (simulated subjects) for each task, to show a substantial effect size of results. Obviously, here and elsewhere p-values improved (but not the effect sizes) when running more simulated subjects.

Simulation Results and Discussion

The RML performance in terms of optimal choice percentages was: Stat = 66.5% (± 4% s.e.m.), Vol = 63.6% (± 1.4% s.e.m.). For Stat2 condition there was no optimal choice, as both options led to reward in 60% of times. Importantly, the model successfully distinguished not only between Stat and Vol environments, but also between Stat2 and Vol, increasing the learning rate exclusively in the latter (Figure 4d). There was a main effect of volatility on learning rate ($F(2,11) = 29$, $p < 0.0001$). Post-hoc analysis showed that stationary conditions did not differ (Stat2 > Stat, $t(11) = 1.65$, $p = 0.13$), while in volatile condition learning rate was higher than in stationary conditions (Vol > Stat2, $t(11) = 5.54$, $p < 0.0001$; Vol > Stat, $t(11) = 5.76$, $p < 0.0001$). Hence, interaction between dACC and LC allows disentangling uncertainty due to noise from uncertainty due to actual changes (Yu and Dayan, 2005; Silvetti et al., 2013a), promoting flexibility (high learning rate) when new information must be acquired, and stability (low learning rate) when acquired information must be protected from noise. This mechanism controls learning rates in both the $dACC_{Action}$ and the $dACC_{Boost}$ modules, thus influencing the whole RML dynamics. The same learning rate effect was found in experimental data (Figure 4e). Indeed, humans increased both learning rate and LC activity only in Vol environments (Silvetti et al., 2013). Thus, humans could distinguish between outcome variance due to noise (Stat2 environment) and outcome variance due to actual environmental change (Vol environment).

The model was also consistent with the fMRI data cited above (Silvetti et al., see Figure 5a). During a RL task executed in the same three statistical environments used in this simulation, the human dACC activity did not follow the pattern in Figure 4d but instead peaked for Stat2 environment, suggesting that activity of human dACC is dominated by prediction error operations rather than by explicit estimation of environmental volatility. Finally, it is worth noting that Vol-related activity of both model and human dACC was higher than in Stat environment (stationary with low

uncertainty), thus replicating the results of previous fMRI study by Behrens et al. (2007).

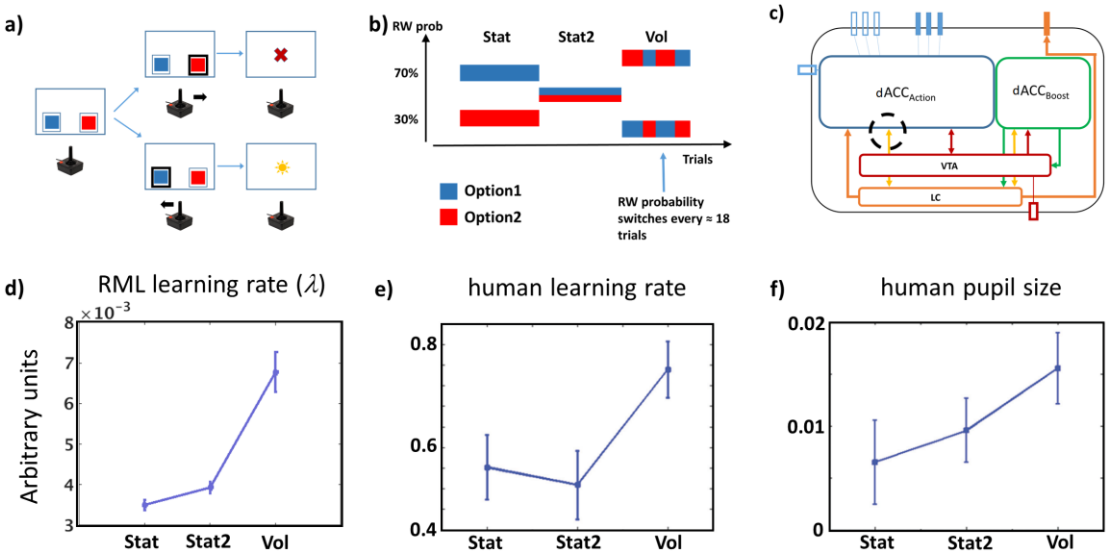

Figure 4. a) The task (2-armed bandit) is represented like a binary choice task (blue or red squares), where the model decisions are represented as joystick movements. After each choice, the model received either a reward (sun) or not (cross). **b)** Example of time line of statistical environments (order of presentation of different environments was randomized across simulations). The plot shows reward probability linked to each option (blue or red) as a function of trial number. In this case the model executed the task first in a stationary environment (Stat), then in a stationary environment with high uncertainty (Stat2), and finally in a volatile (Vol) environment. **c)** Model schema showing where we recorded the signal to measure the learning rate variation (dashed black circle). **d)** Learning rate as a function of environmental volatility (± s.e.m.) in the RML and humans **e)** (modified from: Silvetti et al., 2013a). **f)** human LC activity (inferred by pupil size; Joshi et al. 2016; Varazzani et al. 2015; Aston-Jones and Cohen 2005) during the same task.

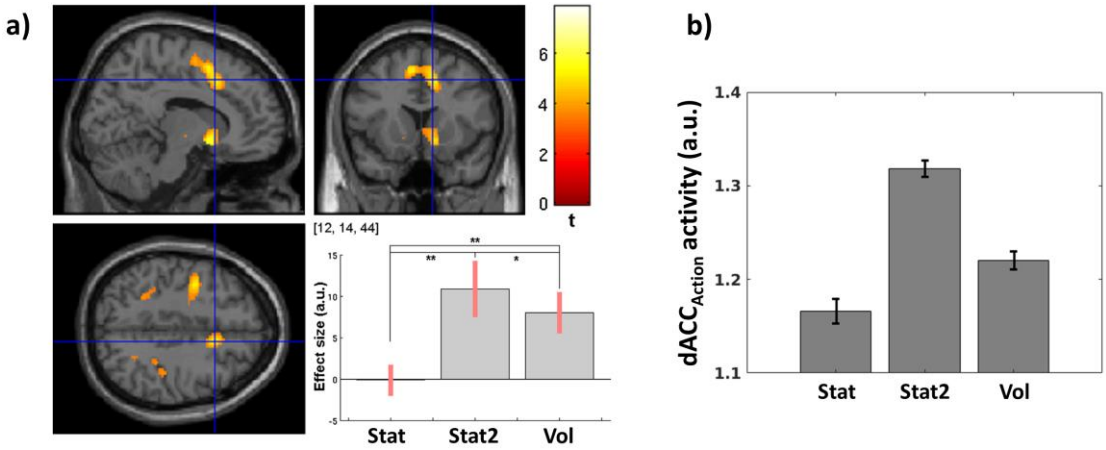

Figure 5. a) dACC activity effect size (extracted from the ROI indicated by cross) in a RL task executed during fMRI scanning. The task was performed in the same three environments we used in our simulations. dACC activity peaked in Stat2 and not in Vol condition (modified from: Silvetti et al., 2013b). **b)** $dACC_{Action}$ average prediction error activity (sum of δ units activity ± s.e.m.) as a function of environmental uncertainty. Differently from the LC, the dACC is maximally active in stationary uncertain environments (Stat2).

Simulation 2: Adaptive physical and cognitive effort control

A long list of experimental results indicates DA and Ne neuromodulators as crucial not only for learning environmental regularities, but also for exerting cognitive control (e.g. Aston-Jones & Cohen 2005; Sara 2009; Vijayraghavan et al. 2007; Langner & Eickhoff 2013; D'Esposito & Postle 2015). Although these mechanisms have been widely studied, only few computational theories explain how the midbrain catecholamine output is controlled to maximize performance (Doya, 2002; Yu and Dayan, 2005), and how the dACC is involved in such a process. In this section, we describe how the $dACC_{Boost}$ module learns to regulate LC and VTA activity to control effort exertion, at both cognitive and physical level (Chong et al., 2017), to maximize long-term reward. In Simulation 2a, we test the cortical-subcortical dynamics regulating catecholamine release in experimental paradigms involving decision-making in physically effortful tasks, where cost/benefit trade –off must be optimized (Salamone et al., 1994). In Simulation 2b, we show how the LC (controlled by the $dACC_{Boost}$) can provide a Ne signal to external "client" systems to optimize cognitive

effort allocation and thus behavioural performance in a visuo-spatial WM task. In both simulations, we also test the RML dynamics and behaviour after DA lesion.

Simulation 2a: Physical effort control and decision-making in challenging cost/benefit trade off conditions

Deciding how much effort to invest to obtain a reward is crucial for human and non-human animals. Animals can choose high effort-high reward options when reward is sufficiently high. The impairment of the DA system strongly disrupts such decision-making (Salamone et al., 1994; Walton et al., 2009). Besides the VTA, experimental data indicate also the dACC as having a pivotal role in decision-making in this domain (Kennerley et al., 2011; Apps and Ramnani, 2014; Vassena et al., 2014). In this simulation, we show how cortical-subcortical interactions between the dACC, VTA and LC drive optimal decision-making when effortful choices leading to large rewards compete with low effort choices leading to smaller rewards, testing whether the RML can account for both behavioral and physiological experimental data. Moreover, we test whether simulated DA depletion in the model can replicate (and explain) the disruption of optimal decision-making, and, finally, how effective behaviour can be restored.

Simulation Methods

We administered to the RML a 2-armed bandit task with one option requiring high effort to obtain a large reward, and one option requiring low effort to obtain a small reward (Walton et al. 2009; here called Effort task; Figure 6a). The task was also administered to a DA lesioned RML.

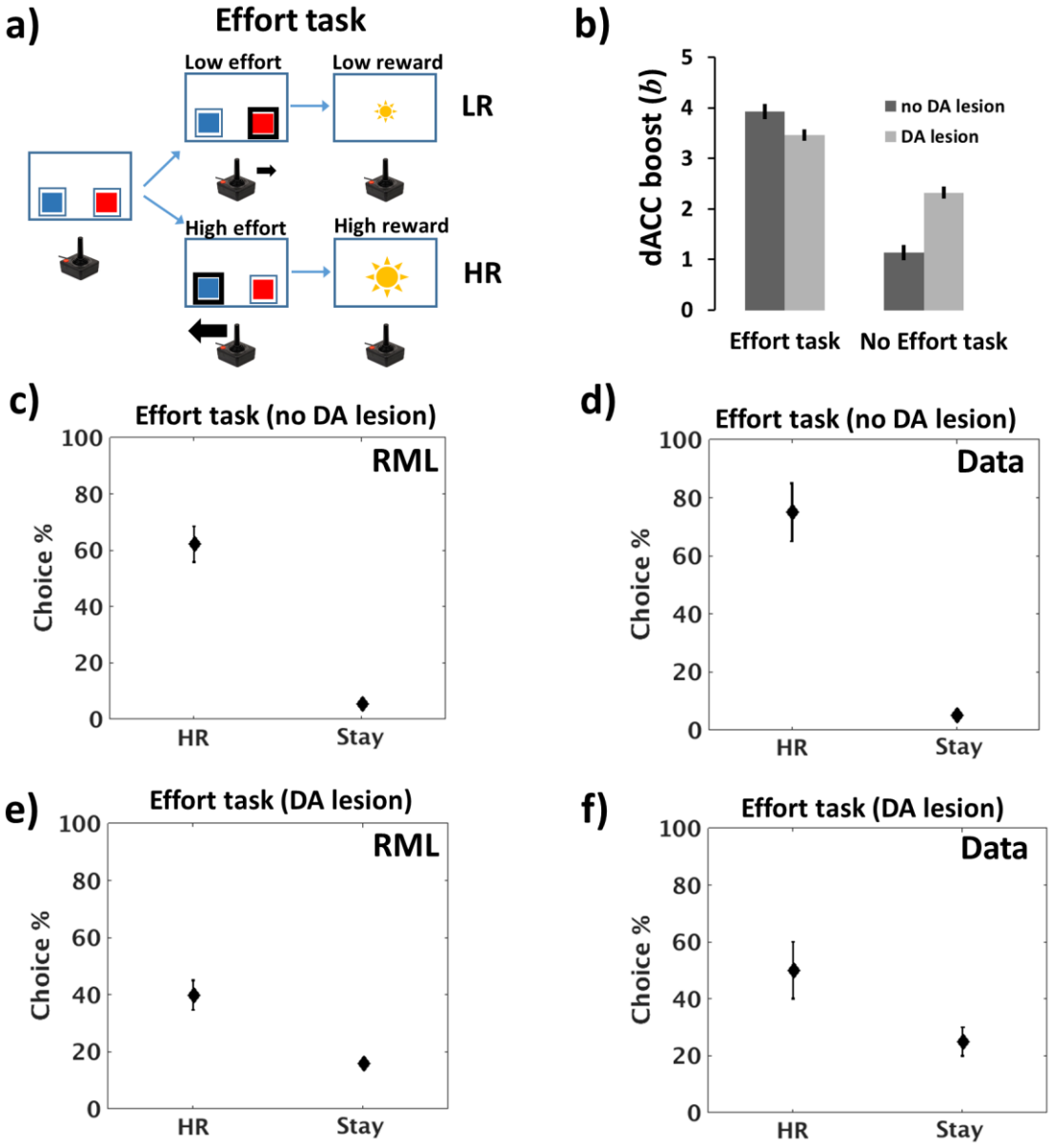

Figure 6. a) Effort task, where a high effort choice (thick arrow from joystick) resulting in high reward (HR, large sun) was in competition with a low effort choice (thin arrow) leading to low reward (LR, small sun). **b)** Catecholamines boosting (*b*) as a function of task type (Effort or No Effort ask) and DA lesion. The boosting value (recorded from the decision units within the $dACC_{Boost}$ module) is higher in the Effort task (main effect of task), but there is also a task x lesion interaction indicating the $dACC_{Boost}$ attempts to compensate the loss of DA, to achieve at least LR (see main text). **c)** Behavioural results (average HR/(LR+HR) ratio ±s.e.m., and average Stay-to-total choices ratio percentage ±s.e.m.) from RML and **d)** empirical data. **e)** Behavioural results after DA lesion in RML and **f)** in empirical data. In this case

animals and the RML switch their preference toward the LR option (requiring low effort). In both d) and f), animal data are from Walton et al. (2009).

Like in Walton et al. (2009), before the execution of the Effort task, the RML learned the reward values in a task where both options implied low effort (No Effort task, Supplementary material, Figure S8a). Besides the high effort and low effort choices, the model could choose to execute no action if it evaluated that no action was worth the reward ("Stay" option).

Simulation Results and Discussion

As shown in Figure 6b, the $dACC_{Boost}$ increased the boosting level (*b* in equations 5, 7) in the Effort task (main effect of task, $F(1,11) = 231.73$, $p < 0.0001$) enhancing both LC and VTA output. The RML learned that boosting is cost/benefit effective when it results in large rewards. Increased Ne affects the Actor (decision-making process), facilitating effortful actions; increased DA affects the learning process of the Critics, increasing the value of the effortful actions. After DA lesion, the $dACC_{Boost}$ decreased the boosting output during the Effort task, while it increased the boosting output during the No Effort task (task x lesion interaction $F(1,11) = 249.26$, $p < 0.0001$). Increased boosting in No Effort task can be interpreted as a compensatory mechanism ensuring a minimal catecholamines level to achieve at least small rewards in the low effort condition, instead of a complete refusal to execute the task (selection of "Stay" option in most of trials).

At behavioural level, in the Effort task, the RML preferred the high effort option to get a large reward (Figure 6c; $t(11) = 4.71$, $p = 0.0042$). After the DA lesion, the preference toward high effort-high reward choice reversed (Figure 6d; $t(11) = -3.71$, $p = 0.0034$). Both these results closely reproduce animal data (Walton et al., 2009). Furthermore, the percentage of "Stay" choices increased dramatically (compare figures 5c and 5e; $t(11) = 18.2$, $p < 0.0001$). Interestingly, the latter result is also in agreement with animal data and could be interpreted as a simulation of apathy due to low catecholamines level. In agreement with animal data, the DA-lesioned RML performance recovers when a No Effort task is administered after the Effort task

(Simulation S2a, Supplementary Results). The same performance recovery occurs in a task where both options are effortful (Double Effort task, Simulation S2b, Supplementary Results), again in agreement with experimental data (Walton et al., 2009).

Simulation 2b: Cognitive effort control in a WM task

Besides attention allocation, Ne neuromodulation plays a crucial role in WM, improving signal to noise ratio by gain modulation mediated by α2-A adrenoceptors (Aston-Jones and Cohen, 2005; Wang et al., 2007), and low level of Ne transmission leads to WM impairment (Li and Mei, 1994; Li et al., 1999). At the same time, it is a major biological marker of effort exertion (Kahneman, 1973; Varazzani et al., 2015). Besides Ne release by the LC, experimental findings showed that slso dACC activity increases as a function of effort in WM tasks (e.g. in mental arithmetic, (Borst and Anderson, 2013; Vassena et al., 2014). Here we show that the same machinery that allows optimal physical effort exertion (Simulation 2a) may be responsible for optimal catecholamine management to control the activity of other brain areas, thus rooting physical and cognitive effort exertion in a common decision-making mechanism. This is possible because the design of the RML allows easy interfacing with external systems. Stated otherwise, the macro-circuit dACC-midbrain works as a "server" providing control signals to "client" areas to optimize their function.

Simulation Methods

We connected the RML system to a WM system (FROST model; Ashby et al. 2005; DLPFC in Figure 2). Information was exchanged between the two systems through the state/action channels in the $dACC_{Action}$ module and the external LC output. The FROST model was chosen for convenience only; no theoretical assumptions prompted us to use this model specifically. FROST is a recurrent network simulating a macro-circuit involving the DLPFC, the parietal cortex and the basal ganglia. This model simulates behavioural and neurophysiological data in several visuo-spatial WM tasks. The external LC output improves the signal-to-noise ratio in the FROST DLPFC. We administered to the RML-FROST circuit a delayed matching-to-sample task with different memory loads (a template of 1, 4 or 6 items to be retained; Figure 7a), running 12 simulations (simulated subjects) for each condition. We used a block

design, where we administered three blocks of 70 trials, each with one specific memory load (1, 4, or 6). In 50% of all trials, the probe fell within the template. The statistical analysis was conducted by a repeated measure 3x2 ANOVA (memory load and DA lesion).

Simulation Results and Discussion

The $dACC_{Boost}$ module dynamically modulate Ne release (*b*, Equation 5) as a function of memory load, in order to optimize performance (Figure 7b, left panel; main effect of memory load on LC output: $F(2,22)= 16.74$, $p < 0.0001$). Like in Simulation 2a, in case of DA lesion, the VTA-dACC-LC interaction is disrupted, leading to a devaluation of boosting and the consequent decision (by the $dACC_{Boost}$ module) of downregulating LC activity (Figure 7c, left panel; main effect of DA lesion on LC output: $F(1,11)= 24.88$, $p < 0.0001$). This happened especially for high memory loads (lesion × memory-load interaction: $F(2,22) = 7.1$, $p = 0.0042$). LC modulation impairment results in poor performance in particular for high memory loads, when high level of Ne is necessary (Figure 7c, accuracy, right panel; lesion × memory-load interaction: $F(2,22) = 8.6$, $p = 0.0017$).

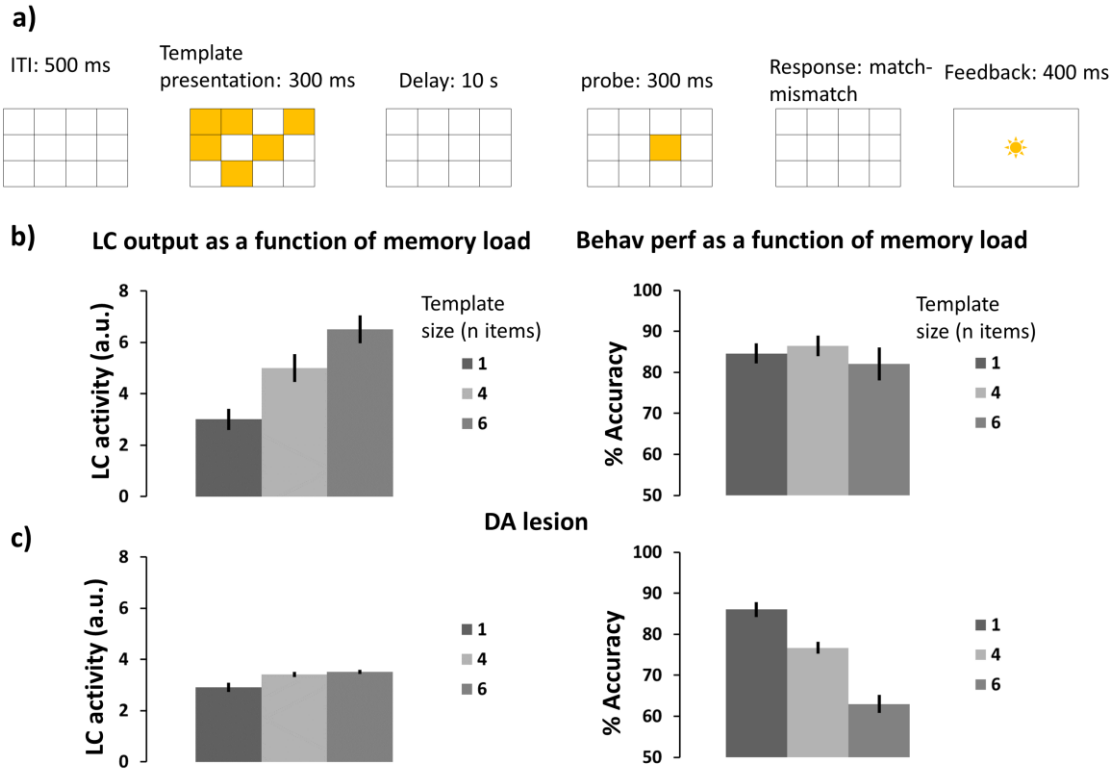

Figure 7. a) Delayed Matching-to-sample task: events occurring in one trial. **b)** Left: LC activity as a function of memory load (number of items presented in the template). Right: behavioural performance as a function of memory load. **c)** LC activity and behavioural performance after DA lesion. Error bars indicate ±s.e.m.

Simulation 3: Reinforcement Learning, meta-learning and higher-order conditioning

Animal behavior in the real world is seldom motivated by conditioned stimuli directly leading to primary rewards. Usually, animals have to navigate through a problem space, selecting actions to come progressively closer to a primary reward. In order to do so, animals likely exploit both model-free and model-based learning (Niv et al., 2006; Pezzulo et al., 2013; Walsh and Anderson, 2014). Nonetheless, model-free learning from non-primary rewards (i.e. higher-order conditioning) remains a basic key feature for fitness, and the simplest computational solution to get adaptive behaviour in complex environments. For this reason, we focus on model-free learning here.

A unifying account explaining behavioral results and underlying neurophysiological dynamics of higher-order conditioning is currently lacking. First, at behavioral level, there is a sharp distinction between higher-order conditioning in instrumental or in classical paradigms. Indeed, although it is possible to train animals to execute complex chains of actions to obtain a reward (instrumental higher-order conditioning, Pierce & Cheney, 2004), it is impossible to install a third- or higher-order level of classical conditioning (i.e. when no action is required to get a reward; O'Reilly et al. 2007). Although the discrepancy has been well known for decades, its reason has not been resolved.

Second, a number of models have considered how TD signals can support conditioning and learning more generally (Holroyd and Coles, 2002; Williams and Dayan, 2005). However, at the best of our knowledge, no TD model addressing DA temporal dynamics also simulated higher-order conditioning at behavioural level.

Here we use the RML to provide a unified theory to account for learning in classical and instrumental conditioning. We show how the RML can closely simulate

the DA shifting in classical conditioning. We also describe how the VTA-dACC interaction allows the model to emancipate itself from primary rewards (higher-order conditioning). Finally, we investigate how the synergy between the VTA-dACC$_{Boost}$ and LC-dACC$_{Boost}$ interactions (the catecholamines boosting dynamics) is necessary for obtaining higher-order instrumental conditioning. This provides a mechanistic theory on why higher-order conditioning is possible only in instrumental and not in classical conditioning.

Simulation 3a: Classical conditioning

A typical experimental finding on DA dynamics is the progressive shifting of DA release onset from primary reward to CS (Schultz et al., 1993). At the same time, omission of expected primary reward typically leads to dips in neural activity in dopaminergic neurons, dropping their discharge rate to zero. DA shifting develops in a spatially discrete manner, i.e. exclusively in the CS-locked and US-locked time windows, without the signal progressively propagating backward from US to CS. We now investigate these properties in the RML.

Simulation Methods

We administered a classical conditioning task, where an environmental cue lasted for 2s, followed by a primary reward on 80% of all trials. Inter trial interval was 4s. The model was trained with 40 trials for each simulation, for each of 12 simulations (subjects).

Simulation Results and Discussion

Figure 8 shows the VTA response (both from RML and animal data) during a classical conditioning paradigm. These results replicate our previous model RVPM simulations (Silvetti et al. 2011). We hypothesized (cf Equation 7) that DA dynamics during conditional learning is determined by dACC-VTA interaction, and more precisely by combining the information from reward expectation and reward PE. This mechanism can closely simulate the progressive shifting of DA activity from reward period to cue period (Figure 8c), and the DA dip when expected rewards are omitted (Figure 8d).

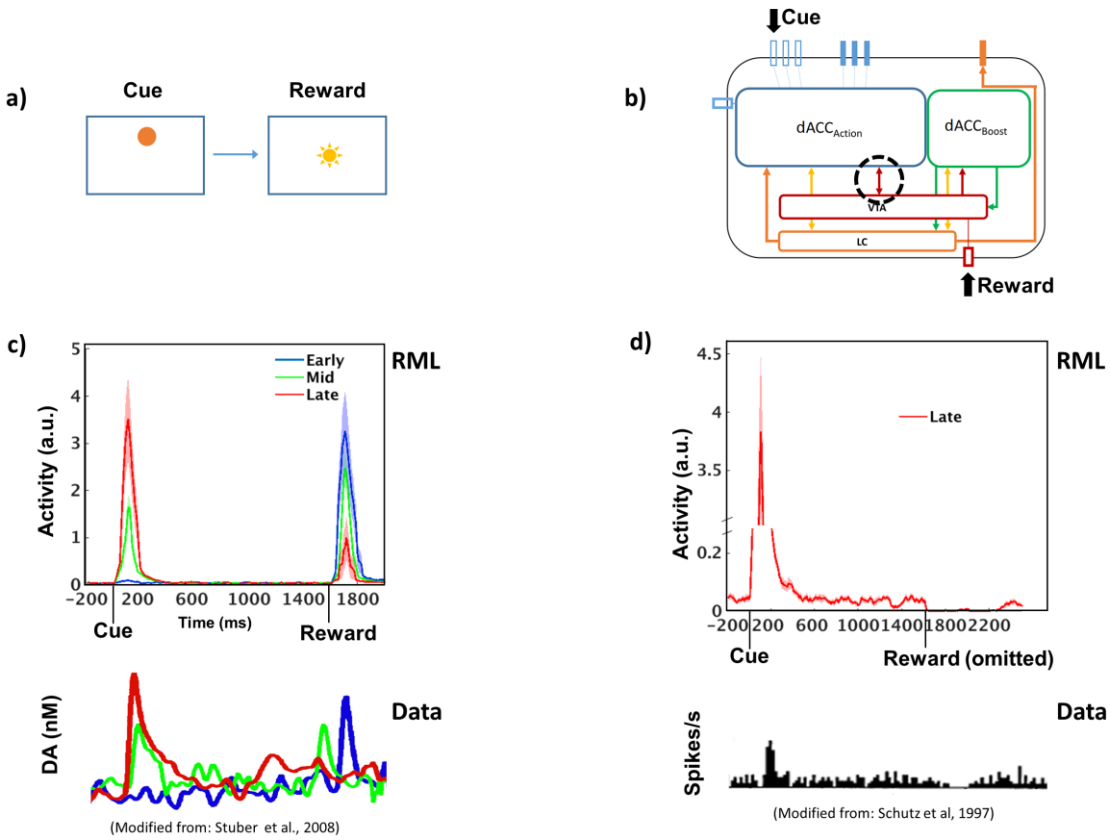

Figure 8. a) Classical conditioning task administered to the model. A cue was presented, then a primary reward was delivered. **b)** Recording site (dashed circle) of VTA activity plotted in c and d (see Supplementary material for more details). Black arrows indicate cue and reward input to the RML. **c)** VTA activity shifting from reward to cue period in three different training phases (early, mid and late). The bottom plot shows empirical data from Stuber et al. (2008). **d)** VTA baseline activity suppression when an expected reward is omitted after extensive conditional training (late training phase). The bottom plot shows empirical data from Schultz et al. (1997).

Simulation 3b: Higher-order classical conditioning

Given the progressive linking of DA response to conditioned stimuli, it is natural to wonder whether a conditioned stimulus can work as a reward itself, allowing to build a chain of progressively higher-order conditioning (i.e. not directly dependent on primary reward). However, classical higher-order conditioning is probably impossible

to obtain in animal paradigms (O'Reilly et al., 2007). We thus investigate what happens in the model in such a paradigm.

Simulation Methods

We first administered the same first-order classical conditioning task of Simulation 3a. We then conditioned a second cue by using the first CS as a non-primary reward. The same procedure was repeated up to third-order conditioning. Each cue was presented for 2s followed by the successive cue or by a primary reward. All cue transitions were deterministic and the reward rate after the third cue was 100%. The reward magnitude was set equal to 7.

Simulation Results and Discussion

In Figure 9 we show the VTA response locked to the onset of each conditioned stimulus. Surprisingly, but in agreement with experimental animal data, the conditioned cue-locked DA release is strongly blunted at the 2nd order, and disappeared almost completely at the 3rd order.

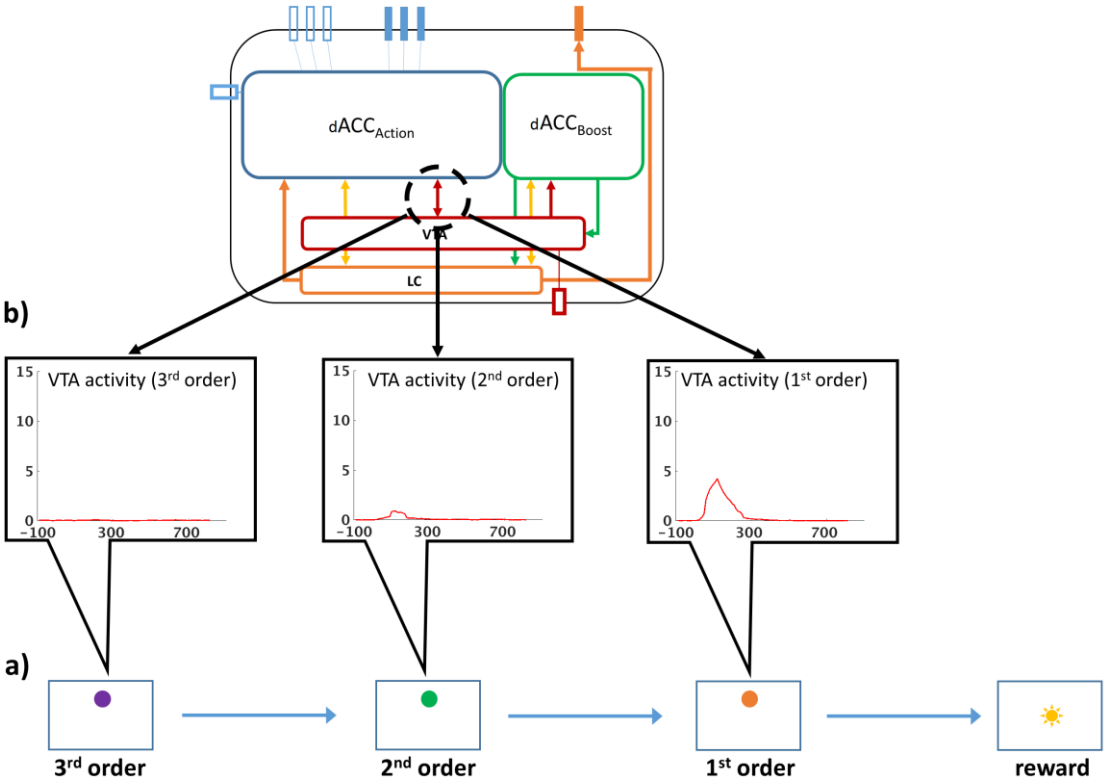

Figure 9. a) Experimental paradigm for higher-order classical conditioning. Sequence of conditioned stimuli (colored disks) followed by primary reward (sun). **b)** VTA activity locked to each of conditioned stimuli. Dashed black circle indicates where the plotted signals were recorded from (see also Supplementary Material).

Simulation 3c: Chaining multiple actions and higher-order conditioning

Differently from classical conditioning paradigms, animal learning studies report that in instrumental conditioning it is possible to train complex action chains using conditioned stimuli (environmental cues) as reward proxies, delivering primary reward only at the end of the task (Pierce and Cheney, 2004).

Simulation Methods

We here administered to the RML a maze-like problem, structured as a series of binary choices before the achievement of a final reward (Figure s6). Each choice led to an environmental change (encoded by a colored disk, like in Figure 3). The training procedure was the same as used for higher-order classical conditioning. We first administered a first-order instrumental conditioning (2-armed bandit task). Then, we used the conditioned environmental cue as non-primary reward to train the RML for second-order conditioning. The procedure was repeated up to third-order conditioning. State-to-state transitions were deterministic and primary reward rate was 100% for correct choices and 0% for wrong choices. Reward magnitude was again set equal to seven.

Simulation Results and Discussion

At the end of training, the system was able to perform three sequential choices before getting a final reward, for an average accuracy of 77.3% (90% C.I. = ±13%) for the first choice (furthest away from primary reward; purple disk, Figure 10a); 95.8% (90% C.I. = [4.2, 5.6]%) for the second; and 98% (90% C.I. = ±0.4%) for the third choice (the one potentially leading to primary reward; orange disk, Figure 10a). Figure 10b shows the cue-locked VTA activity during a correct sequence of choices. Differently from classical conditioning, the DA signal amplitude persists over several

orders of conditioning, making colored disks (also far away from final reward) effective non-primary rewards, able to shape behaviour.

The reason for this difference between classical and instrumental conditioning, is in the role played by the $dACC_{Boost}$ module. This module learns to control the activity of both VTA and LC in order to maximize reward. Boosting catecholamines (Ne and DA) has a cost (Equation 4) and the decision of boosting is selected only when it can result in a substantial performance improvement (in terms of achieved rewards, Equation 4). Figure 10c compares average boosting levels *b* (selected by the $dACC_{Boost}$) in classical and instrumental conditioning. The $dACC_{Boost}$ discovered that boosting Ne and DA was useful in instrumental conditioning; furthermore it discovered that it was not useful in classical conditioning ($t(11) = 5.64$, $p < 0.0001$). This decision amplified DA release during task execution only in instrumental conditioning (compare Figure 10b and Figure 9b). Enhanced VTA activity during the presentation of conditioned stimuli (the colored lights indicating a change in the problem space) means more effective higher-order conditioning, therefore a more efficient behaviour. Conversely, in classical conditioning, the model doesn't need to make any motor decision, as the task consists exclusively of passive observation of incoming cues (colored lights). Therefore, boosting Ne and/or DA does not affect performance (reward amount), as this is completely decided by the environment. In this case, boosting would only be a cost, and the $dACC_{Boost}$ module decides not to boost, with a low VTA activation for conditioned stimuli. This explains the strong limitations in getting higher-order classical conditioning.

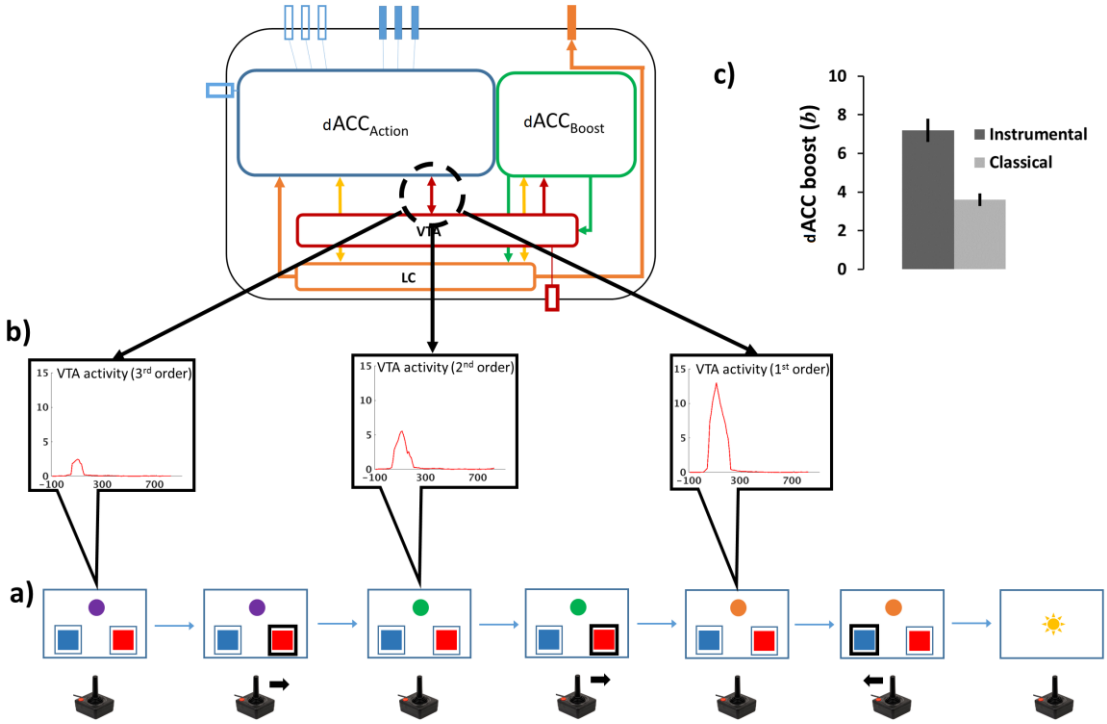

Figure 10. VTA dynamics during higher order instrumental conditioning. **a)** Events occurring during a sequence of correct choices in the task represented also in Figure 3. See Supplementary Methods for details. **b)** Cue-locked (colored disk indicating the environment state) VTA activity. Dashed black circle on the model schema indicates where the plotted signals were recorded from. Differently from higher order classical conditioning, the DA release persists over progressive abstraction of rewards (associative distance from primary reward). **c)** Boosting level (*b*) is higher in instrumental conditioning as compared to classical conditioning (cfr. Figure 9).

General Discussion

We proposed a novel account of the role of dACC in cognition, suggesting that its elusive computational role can be understood by considering its tight interaction with catecholaminergic midbrain nuclei. At a theoretical level, this reconciled three main theoretical frameworks of dACC function, namely behavioural adaptation from a Bayesian perspective (Kolling et al., 2016), effort modulation for cognitive control (Shenhav et al., 2016), and RL-based action-outcome comparison (Silvetti et al., 2014). At an empirical level, the model explained a number of heterogeneous neuroscientific and behavioral data, including error monitoring, learning rate optimization, effort exertion in physical and cognitive tasks, and higher-order classical and instrumental conditioning.

The first meta-learning process we analyzed concerned learning rate (Simulation 1). Earlier *Bayesian models* (e.g. Kalman 1960; Behrens et al. 2007; Mathys et al. 2011) also adapted their learning rates, proposing a computational account of behavioural adaptation. The main limitations of those models are their loose anatomical-functional characterization, the fact that they are computationally hard (in particular for optimal Bayesian solutions, e.g. Behrens et al. 2007), and the need of free parameters specifying a priori the statistical structure of the environment variability (Kalman, 1960; Mathys et al., 2011). The RML instead provides an explicit theory and neuro-computational architecture of how autonomous control of learning rate can emerge from dACC-LC interaction. Moreover, it needed only one minimal and plausible a priori assumption about environmental variability, namely that noise affects environmental signals at much faster scale than volatility.

The second meta-learning process concerned effort exertion and optimal allocation of cognitive and physical resources to achieve a goal (the *cognitive control* perspective; simulations 2a-b). We proposed that investing cognitive and physical effort and controlling the associated costs can be seen as an RL problem (Verguts et al., 2015). Differently from earlier models, the RML generalizes this mechanism to any cognitive domain, showing how the dACC-midbrain system could work as a server providing optimal control signals to other brain areas to maximize success while minimizing costs. Moreover, the RML provides an explicit theory about the role of catecholamines in cognitive control and effort exertion.

The third meta-learning process that we simulated concerned control over *reward* signal for emancipating learning from primary reward (higher-order conditioning; simulations 3b-c). We hypothesized that TD-like DA signals from the VTA are generated thanks to the contribution of afferents from the dACC. Moreover, we modeled DA dynamics at both within and between trials time levels. This approach allowed simulating the DA release shifting in conditioning. At the same time, we showed that stimulus-linked DA signals could play the role of primary reward proxies, to allow reinforcement learning in the absence of primary rewards. Moreover, we showed how dynamic control of DA release by the dACC is necessary for allowing higher-order conditioning, differentiating classical (not effective) vs. instrumental (effective) higher-order conditioning.

Although for sake of simplicity we separated these three domains, in the RML these three meta-learning processes are integrated. Dynamic control of learning rate influences decision-making processes for boosting the release of both Ne and DA, which regulate both effort exertion and DA linking to conditioned stimuli. Finally, DA modulation subserves higher-order conditioning, which allows access to primary rewards in complex tasks, influencing learning rate regulation and boosting levels.

The model also contained two main limitations. First, in our model, DA neurotransmitter plays a role only in learning. Many experimental results suggest that DA is involved not only in learning but also in performance (e.g. attention and WM) modulation (Wang et al., 2004; Vijayraghavan et al., 2007; Shiner et al., 2012; Van Opstal et al., 2014). Our model was intended to be minimalist in this respect, demonstrating how the two neuromodulators can influence each other for learning (DA) and performance (Ne). However, we stress that the ACC_{Boost} control mechanism could be easily, and without further assumptions, extended to DA modulation in the mesocortical dopaminergic pathway, for performance control in synergy with Ne.

The second limitation is the separation of the LC functions of learning rate modulation and cognitive control exertion. The cost of this separation between these two functions is outweighed by stable approximate optimal control of learning rate and catecholamines boosting policy. It must be stressed that the ACC_{Boost} module receives the LC signal λ related to learning rate in any case, making the boosting policy adaptive to environmental changes.

Temporal difference learning and the VTA

Many hypotheses were formulated for understanding both DA shifting from reward onset to CS onset and its computational role in learning. Temporal difference (TD) learning algorithms represent an important framework for explaining DA dynamics in conditional learning. First, in computational neuroscience, TD learning has been typically implemented with environmental state × time step conjunctions for input representations (complete serial compound stimulus; Sutton & Barto 1987). This solution allows learning of outcome timing within a trial and it is used to simulate DA dynamics in classical conditioning (Montague et al., 1996; Schultz et al., 1997; Pan et al., 2005). However, this approach suffers from a series of limitations. It lacks biological plausibility about time representation in the brain, as it assumes that time is represented like an infinite series of separated environmental states (Mauk and Buonomano, 2004; Jin et al., 2009; Bueti and Macaluso, 2010); moreover, it would require a dimensionally infinite time-state vector (where a learned value should be assigned to each state) to simulate continuous neuronal and behavioural dynamics of an animal (or robot) in an ecological context (lack of temporal generalization; Ludvig et al. 2012).

Second, TD applications in machine learning typically implement just environmental states as input (that is, without time representations, Sutton & Barto 1998). Although this solution can be efficient for solving complex tasks, it cannot account for the dynamics of VTA neurons either. In particular, when a reward is expected *n* seconds after a CS, a DA dip is typically observed *n* seconds after CS in trials where the reward does not actually occur (Schultz et al., 1993). A model that does not represent time in one way or another, could not possibly account for this finding.

In contrast, our modified temporal difference signal is consistent with these empirical findings. Because the RML (just like its predecessor the RVPM and other models of conditioning (e.g. O'Reilly et al., 2007)) learns and maintains a separate time signal (*T* in equations 1 and 4), it does account for these two time-related empirical phenomena. This timing signal works as a gating mechanism for generating PEs, and it can be expressed by only two parameters to be learned, i.e. the expected onset of environmental outcome and its duration, differently from the complete serial compound stimulus representation that needs a parameter to be learned for each single

time step. This timing signal may reside in basal ganglia (Brown et al., 1999), cerebellum, or hippocampus, but speculation about this issue is beyond the scope of the current paper (but see e.g. Bueti, Bahrami, Walsh, & Rees, 2010; Jin et al., 2009).

The temporal-difference algorithm implemented in the RML is based on the hypothesis that TD-like dynamics in DA neurons is due to a combination of expected value and PE signals from the dACC. In our model, (the temporal derivative of) expected value (*v* activity) is the key factor allowing higher-order conditioning. This signal links the training DA signal to a conditioned stimulus, which becomes a primary reward proxy for conditioning other stimuli or actions. This solution not only allows simulating VTA dynamics but also provides a mechanistic theory about higher-order conditioning (Simulation 3c). Finally, since the contribution of δ units is locked to outcome onset, it plays no role in higher-order conditioning, although it can contribute to DA signals for first order conditioning.

This temporal difference algorithm is consistent with several ACC-VTA data, some of which were already discussed above. Further relevant data are that there is a strong recurrent connectivity between the dACC and the VTA (Devinsky et al., 1995; Margulies et al., 2007), dACC and VTA activity correlates during RL tasks (Behrens et al., 2007; Jessup et al., 2010; Silvetti et al., 2013b), and dACC stimulation causes VTA response (Gariano and Groves, 1988). Moreover, recent research has started to look at VTA neuron dynamics in more detail, showing a wide range of neural behaviour during conditioning, with most neurons showing a certain degree of response shifting from reward to CS (like in TD algorithm), while others being influenced by either reward expectation or PE (Cohen et al., 2012), or even coding for primary rewards (Takikawa et al., 2004). Such variety of responses could reflect the "components" related to reward expectation $[\dot{v}]^+$ and PE $[\dot{\delta}]^+$ of the complete temporal difference signal represented in Equation 7, which models a prototypical neuron shifting its activity from reward to CS onset. Finally, Eshel et al. (2015) demonstrated that VTA neurons subtract expected value (via GABA neurons). This subtraction would be consistent with the subtracted component $-\left[\dot{\delta}^-\right]^+$ in our VTA signal. The corollary of our approach to VTA dynamics modeling is that DA shifting in conditional learning is a byproduct of stochastic dACC-VTA connectivity, and that PE signals are originally generated in the dACC.

Relationship to other models and the central role of RL in dACC function

The RML belongs to a set of computational models suggesting RL as main function of mammalian dACC. Both the RVPM (RML direct predecessor; Silvetti et al., 2011) and the PRO model (Alexander and Brown, 2011) shares with our model the main idea of the dACC as a state-action-outcome predictor. In all these models, PE plays a core role for learning and decision-making. The RML goes beyond this earlier work, however, by implementing meta-learning and higher-order conditioning. The latter capability emerges from the hierarchical organization of Critic sub-modules, each of which learns CS-outcome associations by using TD-like error signals deriving from hierarchically lower (i.e. closer to primary reward prediction) sub-modules.

Hierarchical organization appears also in other recent dACC models, although it acts on different aspects of learning and decision-making. Holroyd and McClure (2015) proposed a three layered hierarchical RL architecture, where the dorsal striatum played a role of action selector, the dACC of task selector and the prelimbic cortex (in rodents) of context (where and when to execute a task) selector. Moreover, each hierarchical layer implements a PE-based cognitive control signal that discounts option selection costs on the lower hierarchical level. Another recent hierarchical RL model is by Alexander and Brown (2015). In this case the hierarchical design is implemented within the dACC and it unfolds in parallel with a hierarchical model of the DLPFC. In this model, PE afferents from hierarchically lower dACC layers work as an outcome proxy to train higher layers (like the RML); at the same time, error predictions formulated at higher layers of DLPFC modulate outcome predictions at lower ones. This architecture successfully learned tasks where information is structured at different abstraction levels (like the 1-2AX task), exploring the RL basis of autonomous control of information access to WM.

Besides hierarchical organization, the RML represents cognitive control as dynamic selection of effort exertion, a mechanism that has been recently studied also by Verguts et al. (2015). In the latter model, effort exertion was dynamically optimized by the dACC as a process of RL-based decision-making, so that effort levels were selected to maximize long-term reward. This solution successfully simulated many experimental results from cognitive control and effort investment.

Summarizing, the RML shares with other recent computational models of dACC the conceptualization of PE as the core processing operation of this area, the

dACC hierarchical architecture and the capability of dynamic effort exertion. Differently from previous models, the RML integrates all these perspectives, within approximated Bayesian framework, and providing at the same time an explicit theory on how meta-learning could be biologically implemented by the dialogue between the dACC and the midbrain catecholaminergic nuclei LC and VTA. Finally, the neural units composing our model are designed as stochastic leaky integrators, making the RML able to function in continuous time and in presence of noise. These features are crucial to make a model survive outside the simplified environment of trial-level simulations, and make possible to simulate behaviour in the real world, like, for example, in robotics applications.

Future perspectives and experimental predictions

The RML shows how meta-learning involving three interconnected neuro-cognitive domains can account for the flexibility of the mammalian brain. However, our model is not meant to cover all aspects of meta-learning. Many other parameters may be meta-learned too. One obvious candidate is the temperature parameter of the softmax decision process (Khamassi et al., 2015). We recently proposed that this parameter is similarly meta-learned trading off effort costs versus rewards (Verguts et al., 2015). Other parameters from the classical RL modeling include discounting rate or eligibility traces (Schweighofer and Doya, 2003); future work should investigate the computational and biological underpinnings of their optimization.

Given the exceptionally extended dACC connectivity (Devinsky et al., 1995), other brain areas are likely relevant for the implementation of decision making in more complex settings. For example, we only considered model-free dynamics in RL and decision-making. However, both humans and nonhuman animals can rely also on complex environment models to improve learning and decision making (e.g. spatial maps for navigation or declarative rules about environment features). In this respect, future work should particularly focus on dACC-DLPFC-hippocampus interactions (Womelsdorf et al., 2014; Stoll et al., 2016), in order to investigate how environment models can modulate reward expectations and how the nervous system can represent and learn decision tree navigation (Pfeiffer and Foster, 2013).

Given the flexibility of RML and the explicit neurophysiological hypotheses on which it is based, it allows several experimental predictions. For example, negative PE signals are coded by dACC neurons with much higher resolution, as DA neurons

can encode negative PE only by suppressing their baseline activity (Rushworth and Behrens, 2008). This feature, together with the prominent role of the dACC in reward-based decision-making (Rushworth and Behrens, 2008) suggests that PE signals originate from the dACC and are transmitted to the VTA. At experimental level, this hypothesis could be easily tested as it predicts a disruption of DA dynamics in conditional learning after dACC lesion.

A second neurophysiological prediction is about the mechanisms subtending higher-order conditioning and the difference between classical and instrumental paradigms. In the RML, higher-order conditioning is possible only when the agent plays an active role in learning (i.e., instrumental conditioning). We predict that hijacking the dACC decision of boosting catecholamines (e.g., via optogenetic intervention) would make possible higher-order conditioning in classical paradigms (ref. simulations 4-5).

Furthermore, the model provides a promising platform for investigating the pathogenesis of several psychiatric disorders. In a previous computational work, we proposed how motivational and decision-making problems in attention-deficit/hyperactivity disorder (ADHD) could originate from disrupted DA signals to the dACC (Silvetti et al., 2013c). In the current paper, we also simulated a deficit related to cognitive effort (Simulation 3) in case of DA deficit. Together, these findings suggest how DA deficit can cause both motivational and cognitive impairment in ADHD, with an explicit prediction on how DA deficit can impair also Ne dynamics (Hauser et al., 2016) in ADHD. This prediction could be tested by measuring performance and LC activation during decision-making or working memory tasks, while specifically modulating DA transmission in both patients (via pharmacological manipulation) and RML.

Another result with potential translational implication comes from Simulation 2 (and 2b in Supplementary Results), where the model suggested a possible mechanism linking boosting disruption and catecholamines dysregulation. This could be suggestive of pathogenesis of some depressive symptoms. More specifically, the RML predicts that DA antagonization intensifies effort in easy tasks (making them de facto subjectively harder) and decreases it in harder tasks (simulating apathy when effort is required by the environment; Figure 4b). Furthermore, it predicts an increased probability to refuse executing the task (thus simulating apathy). This effect could be experimentally tested by comparing effort-related dACC activation and

behavioral patterns in tasks implying high and low effort with or without DA impairment. Another clinical application concerns a recent theory on autism spectrum disorder (ASD) pathogenesis. (Van de Cruys et al., 2014) proposed that a substantial number of ASD symptoms could be explained by dysfunctional control of learning rate and chronically elevate Ne release. This qualitative hypothesis could be easily implemented and explored quantitatively by altering learning rate meta-learning mechanisms in the RML leading to chronically high learning rate and LC activation.

Summing up, we formulated a model of how dACC-midbrain interactions may implement meta-learning in a broad variety of tasks. Besides understanding extant data and providing novel predictions, it holds the promise of taking cognitive control and, more in general, adaptive behaviour out of the experimental psychology lab and into the real world. In this regard, this work can potentially inform not only what the dACC is doing, but also how to design machine learning algorithms exhibiting mammalian-like cognitive flexibility.

References

Alexander WH, Brown JW (2011) Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci 14:1338–1344.

Alexander WH, Brown JW (2015) Hierarchical Error Representation: A Computational Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex. Neural Comput 27:2354–2410.

Apps MAJ, Ramnani N (2014) The Anterior Cingulate Gyrus Signals the Net Value of Others' Rewards. J Neurosci 34:6190–6200.

Ashby FG, Ell SW, Valentin V V., Casale MB (2005) FROST: A Distributed Neurocomputational Model of Working Memory Maintenance. J Cogn Neurosci 17:1728–1743.

Aston-Jones G, Cohen JD (2005) An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci 28:403–450.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221.

Borst JP, Anderson JR (2013) Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. Proc Natl Acad Sci 110:1628–1633.

Brown J, Bullock D, Grossberg S (1999) How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. J Neurosci 19:10502–10511.

Bueti D, Bahrami B, Walsh V, Rees G (2010) Encoding of temporal probabilities in the human brain. J Neurosci 30:4343–4352.

Bueti D, Macaluso E (2010) Auditory temporal expectations modulate activity in visual cortex. Neuroimage 51:1168–1183.

Chong TT-J, Apps M, Giehl K, Sillence A, Grima LL, Husain M (2017) Neurocomputational mechanisms underlying subjective valuation of effort costs Seymour B, ed. PLOS Biol 15:e1002598.

Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N (2012) Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482:85–88.

Croxson PL, Walton ME, O'Reilly JX, Behrens TEJ, Rushworth MFS (2009) Effort-based cost-benefit valuation and the human brain. J Neurosci 29:4531–4541.

D'Esposito M, Postle BR (2015) The Cognitive Neuroscience of Working Memory. Annu Rev Psychol 66:115–142.

Devinsky O, Morrell MJ, Vogt BA (1995) Contributions of anterior cingulate cortex to behaviour. Brain 118 (Pt 1:279–306.

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

Ebitz RB, Hayden BY (2016) Dorsal anterior cingulate: a Rorschach test for cognitive neuroscience. Nat Neurosci 19:1278–1279.

Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N (2015) Arithmetic and local circuitry underlying dopamine prediction errors. Nature 525:243–246.

Frank MJ, Seeberger LC, O'Reilly R C (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science (80-) 306:1940–1943.

Gariano RF, Groves PM (1988) Burst firing induced in midbrain dopamine neurons by stimulation of the medial prefrontal and anterior cingulate cortices. Brain Res 462:194–198.

Grossberg S (1980) How does a brain build a cognitive code? Psychol Rev 87:1–51.

Hauser TU, Fiore VG, Moutoussis M, Dolan RJ (2016) Computational Psychiatry of ADHD: Neural Gain Impairments across Marrian Levels of Analysis. Trends Neurosci 39:63–73.

Holroyd CB, Coles MG (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev 109:679–709.

Holroyd CB, McClure SM (2015) Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. Psychol Rev 122:54–83.

Jessup RK, Busemeyer JR, Brown JW (2010) Error effects in anterior cingulate cortex reverse when error likelihood is high. J Neurosci 30:3467–3472.

Jin DZ, Fujii N, Graybiel AM (2009) Neural representation of time in cortico-basal ganglia circuits. Proc Natl Acad Sci 106:19156–19161.

Joshi S, Li Y, Kalwani RM, Gold JI (2016) Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. Neuron 89:221–234.

Kahneman D (1973) Attention and effort. Prentice-Hall.

Kalman R (1960) A new approach to linear filtering and prediction problems. J basic

Eng.

Kennerley SW, Behrens TE, Wallis JD (2011) Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat Neurosci 14:1581–1589.

Khamassi M, Quilodran R, Enel P, Dominey PF, Procyk E (2015) Behavioral Regulation and the Modulation of Information Coding in the Lateral Prefrontal and Cingulate Cortex. Cereb Cortex 25:3197–3218.

Kolling N, Wittmann MK, Behrens TEJ, Boorman ED, Mars RB, Rushworth MFS (2016) Value, search, persistence and model updating in anterior cingulate cortex. Nat Neurosci 19:1280–1285.

Kool W, Botvinick M (2013) The intrinsic cost of cognitive control. Behav Brain Sci 36:661–698.

Kool W, McGuire JT, Rosen ZB, Botvinick MM (2010) Decision making and the avoidance of cognitive demand. J Exp Psychol Gen 139:665–682.

Langner R, Eickhoff SB (2013) Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. Psychol Bull 139:870–900.

Li BM, Mao ZM, Wang M, Mei ZT (1999) Alpha-2 adrenergic modulation of prefrontal cortical neuronal activity related to spatial working memory in monkeys. Neuropsychopharmacology 21:601–610.

Li BM, Mei ZT (1994) Delayed-response deficit induced by local injection of the alpha 2-adrenergic antagonist yohimbine into the dorsolateral prefrontal cortex in young adult monkeys. Behav Neural Biol 62:134–139.

Ljungberg T, Apicella P, Schultz W (1992) Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol 67:145–163.

Ludvig EA, Sutton RS, Kehoe EJ (2012) Evaluating the TD model of classical conditioning. Learn Behav 40:305–319.

Margulies DS, Kelly AMC, Uddin LQ, Biswal BB, Castellanos FX, Milham MP (2007) Mapping the functional connectivity of anterior cingulate cortex. Neuroimage 37:579–588.

Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A Bayesian foundation for individual learning under uncertainty. Front Hum Neurosci 5:39.

Mauk MD, Buonomano D V (2004) The neural basis of temporal processing. Annu Rev Neurosci 27:307–340.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. Nat Neurosci 15:1040–1046.

Niv Y, Joel D, Dayan P (2006) A normative perspective on motivation. Trends Cogn Sci 10:375–381.

O'Reilly RC, Frank MJ, Hazy TE, Watz B (2007) PVLV: the primary value and learned value Pavlovian learning algorithm. Behav Neurosci 121:31–49.

Pan W-X, Schmidt R, Wickens JR, Hyland BI (2005) Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci 25:6235–6242.

Pezzulo G, Rigoli F, Chersi F (2013) The Mixed Instrumental Controller: Using Value of Information to Combine Habitual Choice and Mental Simulation. Front Psychol 4:92.

Pfeiffer BE, Foster DJ (2013) Hippocampal place-cell sequences depict future paths to remembered goals. Nature 497:74–79.

Pierce W, Cheney D (2004) Behavior Analysis and Learning New Jersey: Laurence Erlbaum Associates.

Rushworth MF, Behrens TE (2008) Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11:389–397.

Salamone JD, Cousins MS, Bucher S (1994) Anhedonia or anergia? Effects of haloperidol and nucleus accumbens dopamine depletion on instrumental response selection in a T-maze cost/benefit procedure. Behav Brain Res 65:221–229.

Sara SJ (2009) The locus coeruleus and noradrenergic modulation of cognition. Nat Rev Neurosci 10:211–223.

Schultz W, Apicella P, Ljungberg T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci 13:900–913.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science (80-) 275:1593–1599.

Schweighofer N, Doya K (2003) Meta-learning in reinforcement learning. Neural

Netw 16:5–9.

Shenhav A, Botvinick MM, Cohen JD (2013) The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79:217–240.

Shenhav A, Cohen JD, Botvinick MM (2016) Dorsal anterior cingulate cortex and the value of control. Nat Neurosci 19:1286–1291.

Shiner T, Seymour B, Wunderlich K, Hill C, Bhatia KP, Dayan P, Dolan RJ (2012) Dopamine and performance in a reinforcement learning task: evidence from Parkinson's disease. Brain 135:1871–1883.

Silvetti M, Alexander W, Verguts T, Brown JW (2014) From conflict management to reward-based decision making: Actors and critics in primate medial frontal cortex. Neurosci Biobehav Rev 46:44–57.

Silvetti M, Seurinck R, van Bochove ME, Verguts T (2013a) The influence of the noradrenergic system on optimal control of neural plasticity. Front Behav Neurosci in press:160.

Silvetti M, Seurinck R, Verguts T (2011) Value and prediction error in medial frontal cortex: integrating the single-unit and systems levels of analysis. Front Hum Neurosci 5:75.

Silvetti M, Seurinck R, Verguts T (2013b) Value and prediction error estimation account for volatility effects in ACC: A model-based fMRI study. Cortex.

Silvetti M, Wiersema JR, Sonuga-Barke E, Verguts T (2013c) Deficient reinforcement learning in medial frontal cortex as a model of dopamine-related motivational deficits in ADHD. Neural Netw 46:199–209.

Stoll FM, Fontanier V, Procyk E (2016) Specific frontal neural dynamics contribute to decisions to check. Nat Commun 7:11990.

Stuber GD, Klanker M, de Ridder B, Bowers MS, Joosten RN, Feenstra MG, Bonci A (2008) Reward-predictive cues enhance excitatory synaptic strength onto midbrain dopamine neurons. Science (80-) 321:1690–1692.

Sutton R, Barto A (1987) A temporal-difference model of classical conditioning. Proc ninth Annu Conf.

Sutton RS, Barto AG (1998) Reinforcement learning : an introduction. Cambridge (MA): MIT Press.

Takikawa Y, Kawagoe R, Hikosaka O (2004) A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. J Neurophysiol 92:2520–2529.

Van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, de-Wit L, Wagemans J (2014) Precise minds in uncertain worlds: Predictive coding in autism. Psychol Rev 121:649–675.

Van Opstal F, Van Laeken N, Verguts T, van Dijck J-P, De Vos F, Goethals I, Fias W (2014) Correlation between individual differences in striatal dopamine and in visual consciousness. Curr Biol 24:R265–R266.

Varazzani C, San-Galli A, Gilardeau S, Bouret S (2015) Noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. J Neurosci 35:7866–7877.

Vassena E, Silvetti M, Boehler CN, Achten E, Fias W, Verguts T (2014) Overlapping Neural Systems Represent Cognitive Effort and Reward Anticipation Maurits NM, ed. PLoS One 9:e91008.

Verguts T, Vassena E, Silvetti M (2015) Adaptive effort investment in cognitive and physical tasks: a neurocomputational model. Front Behav Neurosci 9:57.

Vijayraghavan S, Wang M, Birnbaum SG, Williams G V, Arnsten AF (2007) Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. Nat Neurosci 10:376–384.

Walsh MM, Anderson JR (2014) Navigating complex decision spaces: Problems and paradigms in sequential choice. Psychol Bull 140:466–486.

Walton ME, Groves J, Jennings KA, Croxson PL, Sharp T, Rushworth MFS, Bannerman DM (2009) Comparing the role of the anterior cingulate cortex and 6-hydroxydopamine nucleus accumbens lesions on operant effort-based decision making. Eur J Neurosci 29:1678–1691.

Wang M, Ramos BP, Paspalas CD, Shu Y, Simen A, Duque A, Vijayraghavan S, Brennan A, Dudley A, Nou E, Mazer JA, McCormick DA, Arnsten AFT (2007) α2A-Adrenoceptors Strengthen Working Memory Networks by Inhibiting cAMP-HCN Channel Signaling in Prefrontal Cortex. Cell 129:397–410.

Wang M, Vijayraghavan S, Goldman-Rakic PS (2004) Selective D2 receptor actions on the functional circuitry of working memory. Science (80-) 303:853–856.

Welch G, Bishop G (1995) An introduction to the Kalman filter.

Williams J, Dayan P (2005) Dopamine, learning, and impulsivity: a biological account of attention-deficit/hyperactivity disorder. J Child Adolesc Psychopharmacol 15:160–169.

Womelsdorf T, Ardid S, Everling S, Valiante TA (2014) Burst firing synchronizes

prefrontal and anterior cingulate cortex during attentional control. Curr Biol 24:2613–2621.

Yu AJ (2007) Adaptive Behavior: Humans Act as Bayesian Learners. Curr Biol 17:R977–R980.

Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. Neuron 46:681–692.