bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

1 Ghent University, Department of Experimental Psychology

2 Donders Institute for Brain, Cognition and Behaviour

* Corresponding author:

massimo.silvetti@ugent.be

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

Acknowledgements.

Conflicts of interest: None. Thanks are due to Clay Holroyd, Gianluca Baldassarre, Daniele Caligiore, Giovanni Pezzulo, and Domenico Maisto for useful comments on this project.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

The dorsal anterior cingulate cortex (dACC) is central in higher-order cognition and behavioural flexibility. Reinforcement Learning, Bayesian decision-making, and cognitive control are currently the three main theoretical frameworks within which the elusive computational nature of this brain area is chased after – but with overall limited success. Here we propose a new model – the Reinforcement Meta Learner (RML) – in which we exploit core insights into the anatomical connections of the ACC with two midbrain catecholamine nuclei (VTA and LC). With its dual role of selecting and implementing optimal decisions (via VTA) and learning to control its own learning parameters (via LC), the RML generates an autonomous control system with the ability of learning to solve hierarchical decision problems without having an intrinsic hierarchical structure itself. We discuss how our model accounts for an unprecedented number of empirical findings across various cognitive domains, assimilates various previously proposed ACC computations while respecting biological constraints, and provides theoretical integration at various levels. The theoretical pillars of our work promise a generic template (i.e., recurrent connectivity between cortex and midbrain) with which meta-cognition can be computationally approached.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 proposes that the dACC is a multi-domain estimator of stimulus and action values for the dACC, aimed at maximizing long-term reward. Yet, the function of dACC extends beyond the role of expectation-outcome comparator; it is capable of adaptive control over its internal parameters. For example, from a cognitive control perspective it has been proposed that dACC controls effort exertion (Shenhav et al., 2013, 2014; 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 pertaining to these perspectives (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.

To fill this lacuna, we here propose a novel RL model coined the Reinforcement Meta Learner (RML) by exploiting for the first time insight into the ACC's anatomical connections to both midbrain catecholamine nuclei (the ventral tegmental area, VTA, and the locus coeruleus, LC). It has been proposed that midbrain catecholamine nuclei play a crucial role in meta-learning (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 argue that the apparent heterogeneity of dACC signals can be integrated when jointly considering dACC and

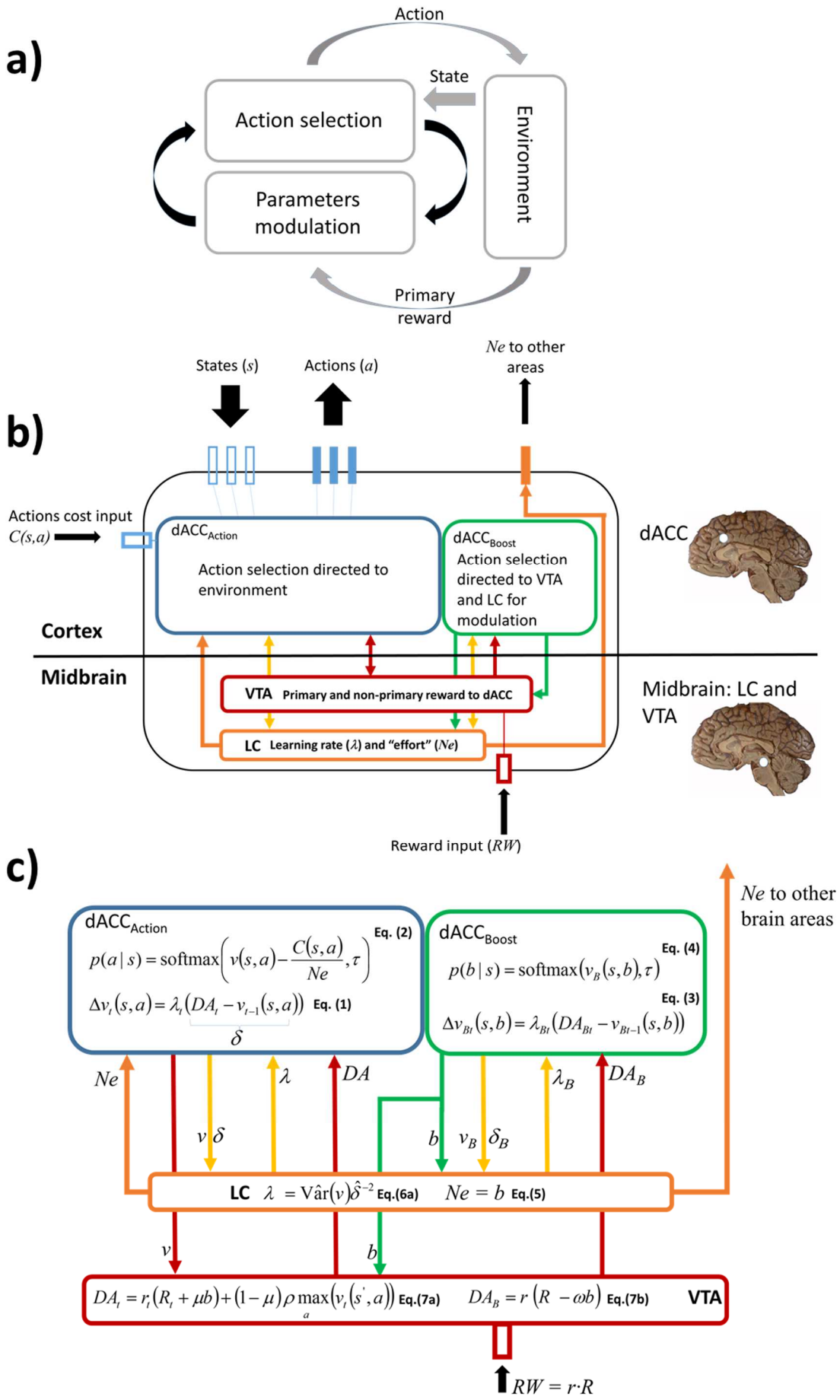
bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the midbrain catecholamine nuclei as a cortical-subcortical loop involved in learning and meta-learning. The RML implements this hypothesis.

The RML goal is to maximize reward not only by making decisions toward the external environment, but also by self-modulating catecholamine release, thus to modify (i.e., meta-learn) its own parameters. We show how this model has sufficient flexibility to generate meta-learning in complete autonomy, by means of the recurrent connectivity between cortex (performing action selection) and midbrain (performing parameters modulation; Figure 1a). Because of the recurrent structure, the model is not hierarchical and performs meta-operations because each circuit modulates the other one. This allows it to solve hierarchical decision problems. The RML can also work as a central "server" of cognitive control, providing control signals to other brain areas to optimize performance. Our model accounts for data across an unprecedented variety of domains, including stability/plasticity balance, effort processing, working memory (WM), higher-order classical and instrumental conditioning.

In the next sections, we first describe in general terms the model structure ('Model Description' section). Second, we outline beforehand how we feel that the RML innovates ACC theory in general (section on 'One framework, several novelties'). Third, we present the results on both neural and behavioural dynamics of the RML in eight (eleven including Supplementary Results) key experimental paradigms, selected on the basis of their popularity and effectiveness in testing RL, cognitive control and Bayesian adaptation phenomena ('Results' section). Critically, the RML captures the relevant findings from these major paradigms without being tuned by experimenters at each simulation. Finally, we wrap up by relating the RML in detail to previous ACC accounts, by describing explicit experimental predictions that derive from the model, and by speculating on the RML's potential application to translational research ('General Discussion' section). Overall, the RML provides for biological grounding, theoretical innovation and integration, and a broad empirical grasp.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 1. a) Model conceptual overview. At the conceptual level, the RML dynamics can be summarized as combining two recurrent circuits. First, an internal loop (black arrows) indicating the recurrent interaction between action-outcome comparison (and action selection) processes and parameter control processes, which together determine the emergence of meta-learning. Second, an external loop (grey arrows), indicating the interactions between agent and environment. **b) Model overview with anatomical-functional analogy.** 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 (*C*), three encoding environmental states (*s*) and one encoding primary rewards (*RW*). The output consists of three channels coding each for one action (*a*), plus one channel conveying LC signals to other brain areas (*Ne*). 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 through first and higher-order conditioning, while $dACC_{Boost}$ modulates catecholamine nuclei output. 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 LC signal controlling costs estimation in the $dACC_{Action}$ is directed also toward other brain areas for neuro-modulation. **c) Model overview with equations.** The equations are reported in their discrete form. Communication between modules is represented by arrows, with corresponding variables near each arrow. Variables δ and δ_B represent the prediction errors from respectively equations 1 and 3. For the complete description of equations and symbols, refer to the Model Description section below.

Model description

Here we describe the general RML principles. The equations are described in a trial-by-trial (discrete-time) form; however, in the simulations, we used a continuous-time implementation. The discrete-time description was chosen to

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

facilitate understanding of the model's conceptual structure; yet, a continuous- time implementation was necessary to study intra-trial dynamics, like DA shifting from US to CS or working memory processing. For a detailed description of the continuous model, refer to Supplementary Methods. To show that the discrete model can reproduce most of the findings obtained with the dynamical model, we ran all the simulations not requiring intra-trial dynamics also with the discrete model, leading to equivalent results (see Supplementary Results: main results replication with discrete model).

The RML approximates optimal decision-making (to maximize long-term reward) based on action-outcome comparisons. Model dynamics can by summarized by Figure 1a, where two loops are described: an inner loop (between action-selection and parameter control processes, black arrows), and an external loop, between RML and environment (blue arrows). The external loop directly influences both parameter control module (through primary rewards) and action selection module (through state transitions), for this reason it can be considered a second communication channel between action selection and parameter control modules. As described more in detail below, both loops contribute to the emergence of meta-learning dynamics.

An overview closer to neurophysiology (Figure 1b-c) shows that action-selection and parameter control belong respectively to cortical and midbrain structures: The dACC (blue and green modules in Figure 1b-c) performs action-outcome comparison and action selection, while it is augmented by meta-learning (LC and VTA; orange and red modules in Figure 1b-c).

We designed the model such that the 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. There are two other input channels, one dedicated to primary reward from environment and the other to signal indicating action costs. Finally, there is one output channel conveying norepinephrine (*Ne*) signals to other brain areas.

In Figure 1c we link the model equations described below with a graphical representation. As we describe in detail in the following section, the RML autonomously modulates three different parameters providing a near optimal solution for the following meta-learning problems. First, modulation of *learning rate* (λ) can

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

make sure knowledge 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 *N*e) 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 *DA*) 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). Moreover, adaptive *DA* signal implies modulating the (perceived) value of state/action couples (*DA* represents outcome in Equation 1), eventually self-motivating the RML in choosing one specific action even if it implies a high execution cost (*C*, Equation 2), and thus influencing the RML behaviour also in challenging benefit-cost problems. Meta-learning processes involving those three signals (learning rate, cost estimation, reward) interact with each other, based on the dynamics described below, and thus forming an integrated system where flexibility emerges from its recurrent dynamics.

The RML is scalable by design, i.e. there is no theoretical limit to the number of state/action channels, and neither the number of parameters nor their values changes as a function of task type/complexity. As specified below, we used a single set of parameters across all simulations both for the discrete model (Table 1) and for the dynamic model (see Table s1).

Parameter	Value	Meaning	Equation
ρ	0.2	TD-learning signal decay	7a
μ	0.3	DA dynamics	7a
τ	0.6	Softmax temperature	2
α	0.3	Kalman filtering meta-parameter	6c-d
β	0.2	Learning rate lower bound	6
ω	0.15	Boosting cost	7b

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1. Parameters list and values for the discrete model. We ran all the simulations using one set of parameter values (for both the discrete and dynamical model).

dACC$_{\text{Action}}$

The dACC$_{\text{Action}}$ module consists of an Actor-Critic system augmented with meta-learning (Figure 1c, blue box). The 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. The Actor selects motor actions (based on Critic expectation) to maximize long-term reward.

The central equation in this module governs Critic state/action value updates:

$$\Delta v_t(s,a) = \lambda_t\left(DA_t - v_{t-1}(s,a)\right) \tag{1}$$

where $v(s,a)$ indicates the value (outcome prediction) of a specific action a given a state s. Equation 1 ensures that v comes to resemble the environmental outcome encoded by dopaminergic signal (DA), which is generated by the VTA module (Figure 1; Equation 7). It entails that the update of v at trial t is based on the difference between prediction (v) and outcome (DA), which defines the concept of PE. The latter is weighted by learning rate λ, making the update more (high λ) or less (low λ) dependent on recent events. We propose that λ itself is modulated by the LC based on v and PE signals from the dACC$_{\text{Action}}$ (Equation 6a).

The DA signal, afferent from the VTA (Equation 7), is either linked to primary or non-primary reward (higher-order conditioning) and is modulated by the dACC$_{\text{Boost}}$ module via parameter b (Equation 7a).

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

$$p(a \mid s) = \text{softmax}\left(v(s,a) - \frac{C(s,a)}{Ne}, \tau\right) \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Function C assigns a cost to each state/action couple, for example energy depletion consequent to climbing an obstacle. C is modulated by norepinephrine afferents from LC (Ne), which is itself controlled by the $dACC_{Boost}$ module, via parameter b. Ne levels discount C, lowering the perceived costs and energizing behaviour. Here we remind the reader that the RML can choose not to engage the task ("Stay"); this option has $C = 0$. In this way, a high level of Ne energizes behaviour, promoting both high cost actions and reducing the probability that the RML chooses to "Stay".

The dynamical form of these equations is described in the *$dACC_{Action}$-VTA system* paragraph in Supplementary Methods.

$dACC_{Boost}$

The $dACC_{Boost}$ module is an Actor-Critic system that learns only from primary rewards (Figure 1c, green box). This module controls the parameters for cost and reward signals in equations 1-2 ($dACC_{Action}$), 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 $dACC_{Boost}$ updates the boost values v_B(s,b), via the equation:

$$\Delta v_{B,t}(s,b) = \lambda_{B,t}\left(DA_{B,t} - v_{B,t-1}(s,b)\right) \tag{3}$$

Equation 3 represents the value update of boosting level b in the environmental state s. The $dACC_{Boost}$ module receives dopaminergic outcome signals (DA_B) from the VTA module. As described in Equation 7b, DA_B represent the reward signal discounted by the cost of boosting catecholamines (Kool et al., 2010; Kool and Botvinick, 2013; Shenhav et al., 2013). Also in Equation 3 there is a dynamic learning rate (λ_B), estimated by Equation 6a in the LC. The Actor submodule selects boosting actions to maximize long-term reward:

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$p(b \mid s) = \mathrm{softmax}\left(v_B(s,b), \tau\right) \tag{4}$$

Referring to Equation 1, the $dACC_{Boost}$ modulates the reward signal by changing the *DA* signal coded in VTA (Equation 7a). Furthermore, $dACC_{Boost}$ also modulates the cost signal by changing parameter *Ne* (via LC module, Equation 5) in the function representing action cost *C* (Equation 2; represented in the Actor within the $dACC_{Action}$). The dynamical form of these equations is described in the *$dACC_{Boost}$-LC-VTA system* paragraph in Supplementary Methods.

LC

LC control over effort exertion and behavioural activation

The LC module plays a double role (Figure 1c, orange box). First it controls cost via parameter *Ne*, as a function of boosting value *b* selected by the $dACC_{Boost}$ module. For sake of simplicity, we assigned to *Ne* the value of *b*; any monotonic function would have played a similar role.

$$Ne = b \tag{5}$$

The *Ne* signal is directed also toward external brain areas as a performance modulation signal (Figure 2; Simulation 2b).

LC control over learning rate

The LC module optimizes also learning rate in the two dACC modules (λ and λ_B). Approximate optimization of 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 the dACC (both modules) and the LC module, which controls learning rate based on the signals afferent from the dACC. 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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 priors, so that high values reflect low confidence in priors and more influence of evidence on posteriors estimation.

The main limitation of this and similar methods is that one must know a priori the model describing the environment statistical properties (noise and process variance). This information is typically inaccessible by biological or artificial agents, which perceive only the current state and outcome signals from the environment. Our LC module bypasses this problem by an approximation based on the information afferent from the dACC, without knowing a priori neither process nor noise variance. To do that, the LC modulates λ (or λ_B) as a function of the ratio between the estimated variance of state/action-value ($\hat{\mathrm{Var}}(v)$) over the estimated squared PE ($\hat{\delta}^2$):

$$\lambda_t = \frac{\hat{\mathrm{Var}}(v)_t}{\hat{\delta}_t^2} \tag{6a}$$

With process variance given by:

$$\hat{\mathrm{Var}}(v) = (v_t - \hat{v}_{t-1})^2 \tag{6b}$$

Where $\hat{v}$ is the estimate of v, obtained by low-pass filtering tuned by meta-parameter α:

$$\hat{v}_t = \hat{v}_{t-1} + \alpha(v_t - \hat{v}_{t-1}) \tag{6c}$$

The same low-pass filter is applied to PE signal (δ) to obtain a running estimation of total variance $\hat{\delta}^2$, which corresponds to the squared estimate of unsigned PE:

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$\hat{\delta}_t = \hat{\delta}_{t-1} + \alpha\left(\left|\delta_t\right| - \hat{\delta}_{t-1}\right) \tag{6d}$$

For numerical stability reasons, λ was constrained in an interval between 1 (max learning rate) and β (min learning rate).

In summary, in equations 6a-d Kalman gain is approximated using 3 components: reward expectation (v), PE signals (δ) (both afferent from the dACC modules) and a meta-parameter (α), defining the low-pass filter to estimate process and total variance. The meta-parameter α represents the minimal assumption that noise-related variability occurs at a faster time scale than volatility-related variability. Equations 6a-d are implemented independently for each of two dACC modules, so that each Critic 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 6a.

The dynamical form of these equations is described in the *dACC-LC system* paragraph in Supplementary Methods.

VTA

The VTA provides training signal *DA* to both dACC modules, either for action selection directed toward the environment (by $dACC_{Action}$) or for boosting-level selection (by $dACC_{Boost}$) directed to the midbrain catecholamine nuclei (Figure 1c, red box). 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).

$$DA_t = r_t\left(R_t + \mu b\right) + \left(1-\mu\right)\rho \max_a\left(v_t\left(s^{'}, a\right)\right) \tag{7a}$$

Equation 7a represents the modulated (by b) reward signal. Here, r is a binary variable indicating the presence of reward signal, and R is a real number variable indicating reward magnitude. Parameter ρ is the TD discount factor, while parameter μ is a scaling factor distributing the modulation b between primary (first term of the

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

equation) and non-primary (second term) reward. It is worth noting that when $\mu = 0$, Equation 7a simplifies to a Q-learning reward signal.

The VTA signal directed toward the $dACC_{Boost}$ is described by the following equation:

$$DA_{Bt} = r_t\left(R_t - \omega b\right) \tag{7b}$$

where ω is a parameter defining the cost of catecholamine boosting (Kool et al., 2010; Kool and Botvinick, 2013; Shenhav et al., 2013). In summary, boosting up DA by b (Equation 7a), can improve behavioural performance (as shown in simulations below) but it also represents a cost (Equation 7b). The $dACC_{Boost}$ module finds the optimal solution for this trade-off, choosing the optimal DA level to maximize performance while minimizing costs (for a formal analysis about this optimization process we refer to Verguts et al., 2015). The dynamical form of these equations is described in the *$dACC_{Action}$-VTA* and *$dACC_{Boost}$-VTA* paragraphs in Supplementary Methods.

Client/server system

Finally, the RML can optimize performance of other brain areas. It does so via the LC-based control signal (*Ne*), which is the same signal that modulates costs (Equation 2; 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 other areas. Moreover, because optimization of any brain area improves behavioural performance, the $dACC_{Boost}$ can modulate (via LC signals) any cortical area to improve performance (see Simulation 2c)

Example of RML dynamics

To visualize the RML functions, we show in Figure 2 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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 four 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 (Ne), and reward signals (b, Equation 7), each of which divided in sub-sections, for a total of eleven experimental results (including three in the Supplementary Results). Across all simulations, the exact same RML architecture and parameter setting was used.

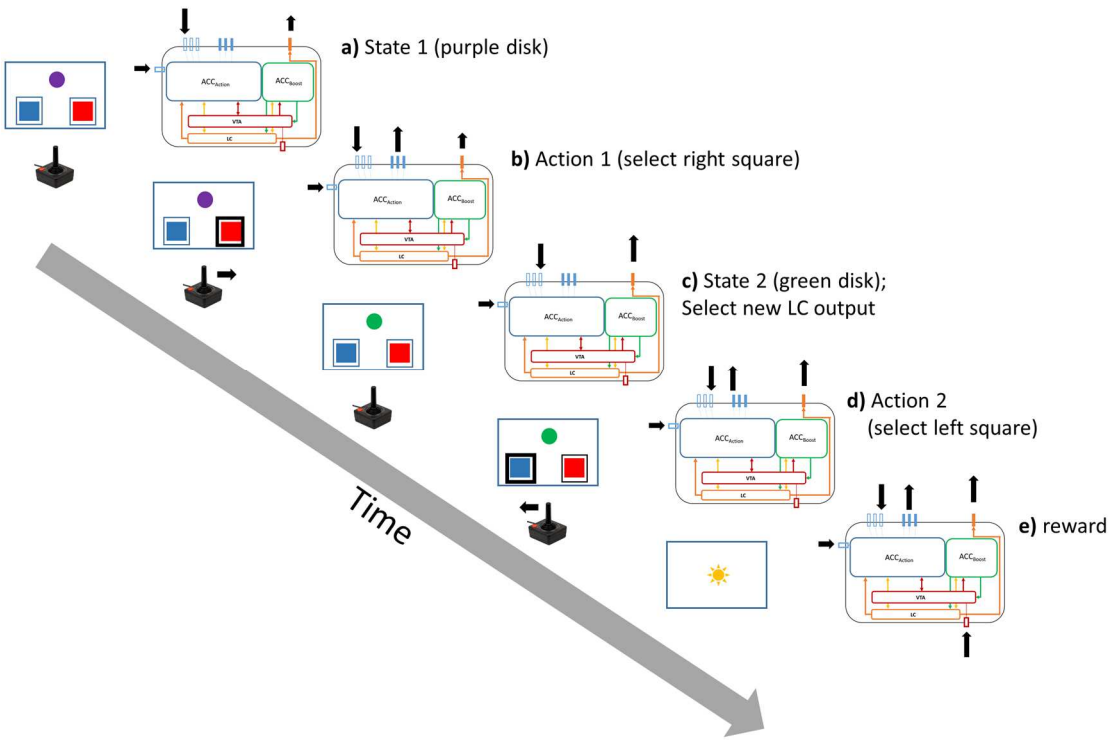


Figure 2. Example of input-output RML dynamics (right sequence) during a trial in a higher-order instrumental conditioning task (left sequence). The task is represented like a 2-armed bandit task, where two options (red and blue squares) can be selected (joystick). The RML needs to select a sequence 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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

One framework, several novelties

Before displaying the results from the RML performance in behavioural tasks, in this paragraph we summarize the computational and neuroscientific novelties of our model. For a comparison with other existing models, see the dedicated paragraph in the General Discussion section.

Meta-learning without hierarchy

One key feature of the RML is that its meta-decision capabilities do not derive from a hierarchical architecture, but instead emerge from a circular, ongoing interaction between cortical and subcortical circuits. For example, selection of appropriate action may be conditional on the brain's conception of the reward environment as either stationary or volatile. Our model solves this apparent hierarchical decision tree via recurrent connectivity between cortex and midbrain – with the dACC sending RL signals to the LC, which then adjusts the learning rate of the dACC. Notably, the loop between cortex and midbrain is in part closed by including also the environment (Figure 1a). For example, $dACC_{Boost}$ module decisions change the activity of both LC and VTA, which change the activity of the $dACC_{Action}$. The dACC_action defines behaviour toward the environment that provides feedback to both the dACC modules, thus closing the loop. The RML needs no homunculus-based regulations and defines

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

autonomously its own activity level (see simulations 2a-b), possibly deciding to reduce behavioural activity if the environment is too frustrating or not rewarding enough (e.g. see Simulation 2a-b).

Bayesian adaptive learning

In the RML, learning rate control emerges from the dialogue between dACC and LC. More precisely, volatility effects occur in the LC (which controls learning rate by processing RL signals afferent from the dACC), while dACC computes state/action-outcome comparisons and generates the RL signals used by the LC to control the learning rate. We consider the dACC as the source of raw data (RL signals) about environment statistical structure, which are processed by the LC to estimate environmental volatility. This hypothesis not only proposes a novel implementation of near optimal learning rate control, but it also unifies RL and Bayesian perspectives, and explains how the mammalian brain can estimate environmental volatility by using RL signals generated during animal-environment interaction (See "LC control over learning rate" in Model Description section above).

Cognitive control with Bayesian flexibility

Bayesian adaptive learning determines the machinery of both the dACC modules in our model. Like the $dACC_{Action}$, the $dACC_{Boost}$ module is a RL-based decision-maker with Bayesian control over learning rate. This means that the modulation of catecholamines release (of which the $dACC_{Boost}$ is a key element) itself is based on Bayesian learning (the dynamics between $dACC_{Boost}$ and LC is described in "LC control over learning rate" paragraph). This aspect provides Bayesian flexibility to cognitive control, a novelty that unifies both these theoretical perspectives.

Reward subjectivity and effort modulation: two sides of the same coin

We investigate also the effects of adaptive modulation on reward signal themselves (primary and non-primary) both at the neurophysiological and behavioural level. VTA modulation by dACC emancipates reward from its connotation of "objective" environmental signal and introduces the component of subjectivity in reward "perception" by the mammalian brain. Our model can autonomously "boost" reward signals from the external environment to improve motivation and maximize behavioural efficiency. Moreover, as VTA activity is modulated by the same signal

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

afferent to the LC (*b* from $dACC_{Boost}$), this feature provides a unified theoretical view between optimal effort allocation and control over motivational and learning aspects, a property that allowed to explain some behavioural and neurophysiological aspect of higher-order conditioning (Simulation 3)

dACC-midbrain circuit as a general provider of control signals to other brain areas. We propose that the dACC-midbrain ensemble works as a system that learns and decides how to influence other brain areas (via *Ne* signals directed toward the environment). In Simulation 2b we tested this hypothesis by connecting the RML to a previously published WM model. At the best of our knowledge, this is the first model that improves the performance of a different, independently designed and published model.

Results

In this section, we report the results obtained with the dynamical model. Simulation results from the discrete model are equivalent and reported in paragraph "Supplementary Results: main results replication with discrete model". Simulation 2c (related to WM task) required intra-trial dynamics, therefore, for that one, we used only the dynamical model.

To mimic standard experimental paradigms as closely as possible, we repeated just 12 times each simulation (simulated subjects), to show a substantial effect size of results. Obviously, p-values improved (but not the effect sizes) when running more simulated subjects.

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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RML implements learning rate meta-learning by means of recurrent interaction between the dACC and the LC (Equation 6a), allowing it to estimate these quantities too, thus to approximate optimal control.

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 yet been reconciled in 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 3a-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 (see: Supplementary Results: dynamical model; Figure s7), 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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 3d). There was a main effect of volatility on learning rate λ ($F(2,11) = 29$, $p < 0.0001$). 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 3e). 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 4). 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 3d 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 (Figure 4a). The RML captures these results (Figure 4b) by its PE activity (activity of δ units from equations s3-4). 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 studies by Behrens et al. (2007).

Summarizing, results in Figure 3d show the effectiveness of our RL-based Kalman approximation, suggesting that the LC codes for volatility and that our algorithm modeling dACC-LC dialogue is both computationally effective and

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

neurophysiologically grounded. Results in Figure 4b, on the other hand, indicate that, although the dACC is part of a Bayesian estimator, its activity is mostly influenced by overall environmental variance, rather than coding for volatility.

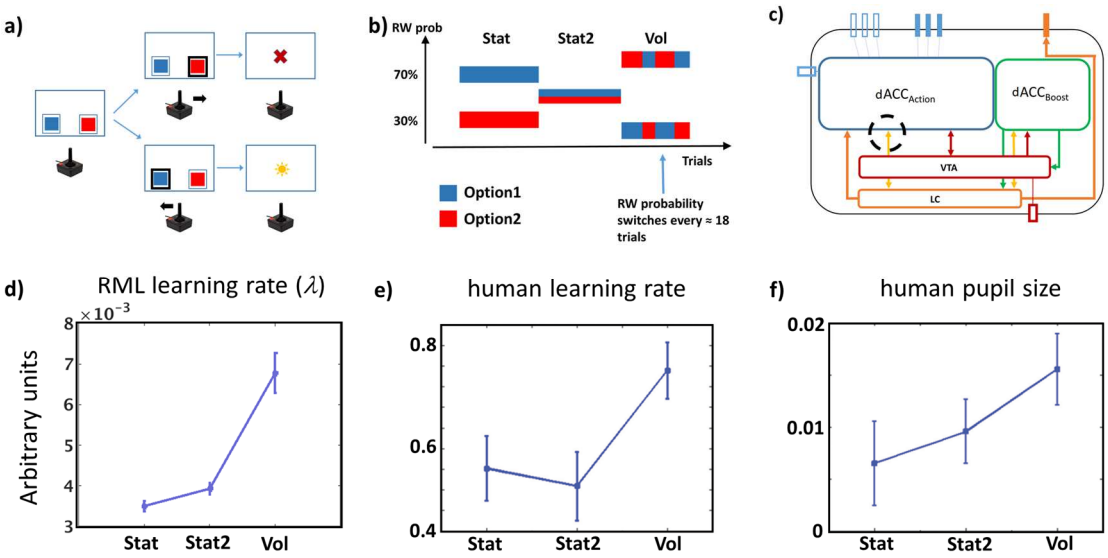


Figure 3. a) The task (2-armed bandit) is represented like a binary choice task (blue or red squares), where the model decisions are represented as joystick movements. After each choice, the model received either a reward (sun) or not (cross). **b)** Example of time line of statistical environments (order 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.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

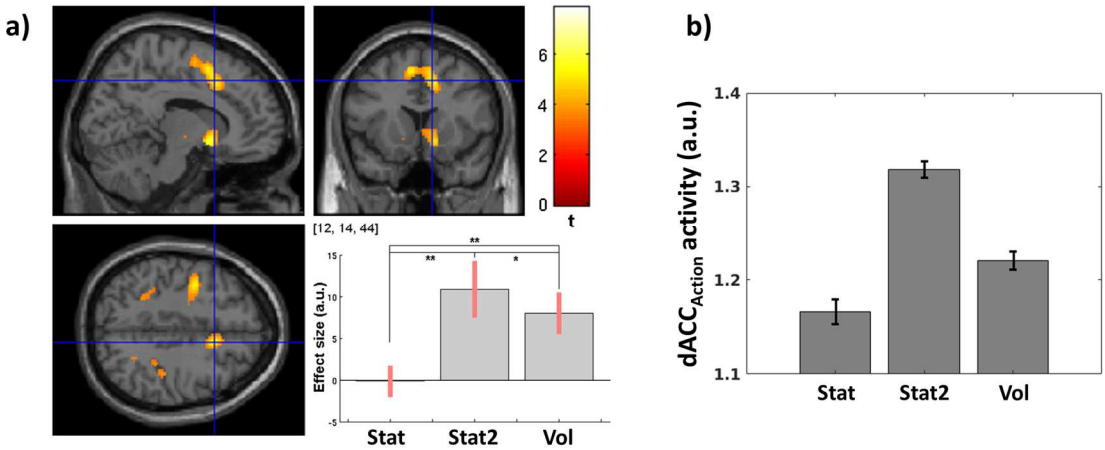


Figure 4. a) dACC activity effect size (extracted from the ROI indicated by cross) in a RL task executed during fMRI scanning. The task was performed in the same three environments we used in our simulations. dACC activity peaked in Stat2 and not in Vol condition (modified from: Silvetti et al., 2013b). **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: Adapting physical and cognitive effort

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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$dACC_{Boost}$) 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 may drive optimal decision-making when effortful choices leading to large rewards compete with low effort choices leading to smaller rewards. We thus test whether the RML can account for both behavioral and physiological experimental data. Moreover, we test whether simulated DA depletion in the model can replicate 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 5a). The task was also administered to a DA lesioned RML (simulated by reducing VTA activity to 60%, see Suppl. Methods).

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

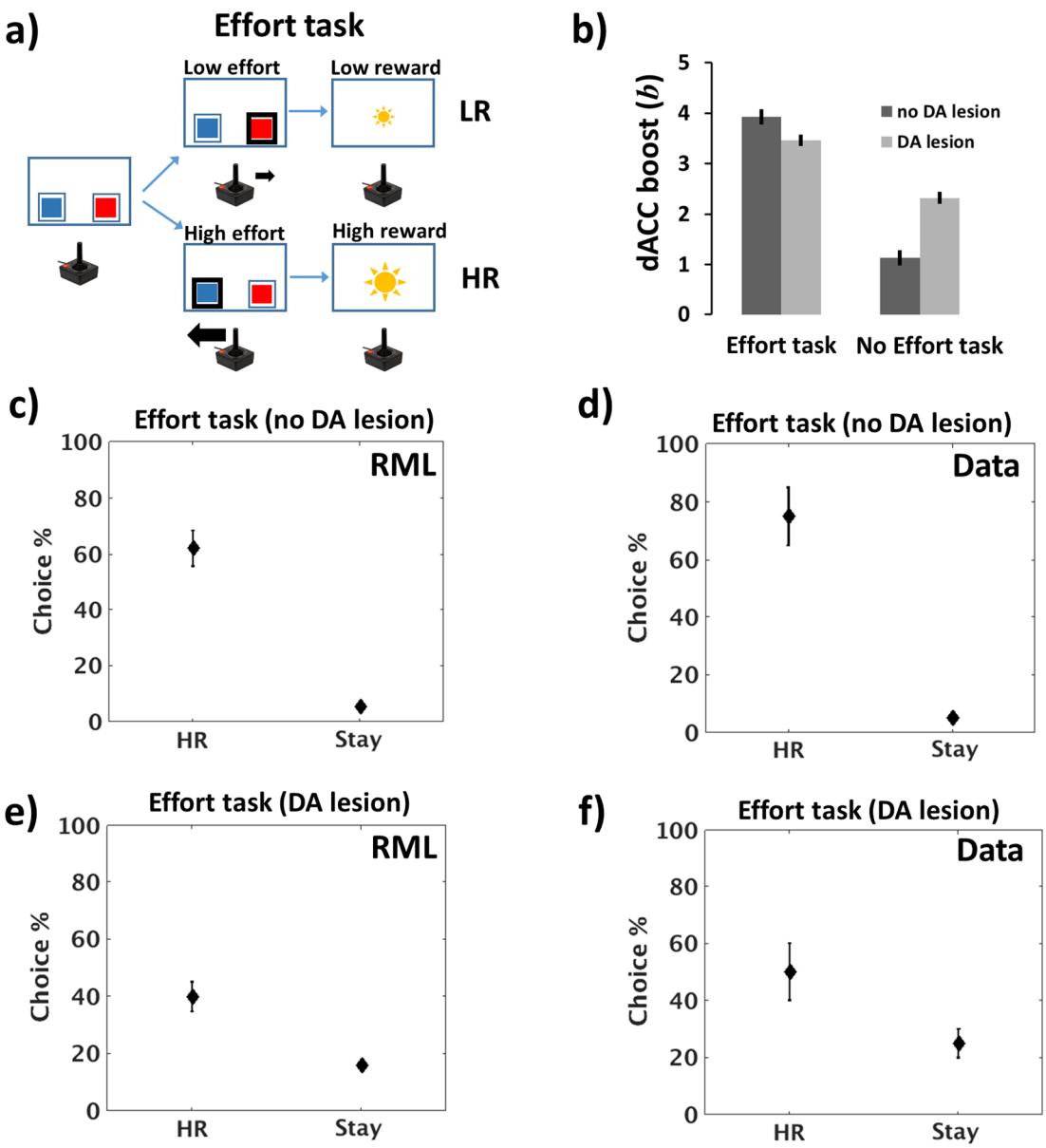


Figure 5. a) Effort task, where a high effort choice (thick arrow from joystick) resulting in high reward (HR, large sun) was in competition with a low effort choice (thin arrow) leading to low reward (LR, small sun). **b)** Catecholamines boosting (*b*) as a function of task type (Effort or No Effort task) 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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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). 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 5b, the $dACC_{Boost}$ increased the boosting level (*b*) in the Effort task (main effect of task, $F(1,11) = 231.73$, $p < 0.0001$) enhancing both LC and VTA output (equations 5, 7a). Increased *Ne* influences the Actor (effect of *Ne* on action cost estimation in decision-making process, Equation 2), facilitating effortful actions, while increased DA affects the learning process of the $dACC_{Action}$ (equations 1, 7a), increasing the reward signal related to effortful actions (therefore "subjectively" increasing their value with respect to the "objective" reward signal provided by the environment). At the same time, the $dACC_{Boost}$ had to face the cost of boosting (ωb term in Equation 7b), so that the higher *b,* the higher was the reward discount for the $dACC_{Boost}$ module. The result of these two opposite forces (maximizing performance by catecholamines boosting and minimizing the cost of boosting itself) converges to the optimal value of *b* and therefore of catecholamines release by VTA and LC (Figure 6a).

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 $F(1,11) = 249.26$, $p < 0.0001$). Decreased boosting derives from decreased DA signal to $dACC_{Boost}$ module (Figure 6b). On the other side, increased boosting *b* in No Effort task can be interpreted as a compensatory mechanism ensuring the minimal catecholamines level to achieve the large reward when just a low effort is necessary (Figure 6d). Indeed, the lack of compensation in No Effort task would result in an apathic policy, where the RML would often select the "Stay" action, saving on minimal costs of moving but also reducing the amount of reward. When the incentive is high (high reward available) and the effort required to obtain

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the reward is low, the RML predicts that the DA lesioned animal would choose to exert some effort (boosting up the remaining catecholamines) to promote active behaviour versus apathy.

At behavioural level, in the Effort task, the RML preferred the high effort option to get a large reward (Figure 5c; $t(11) = 4.71$, $p = 0.0042$). After the DA lesion, the preference toward high effort-high reward choice reversed (Figure 5d; $t(11) = -3.71$, $p = 0.0034$). Both 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.

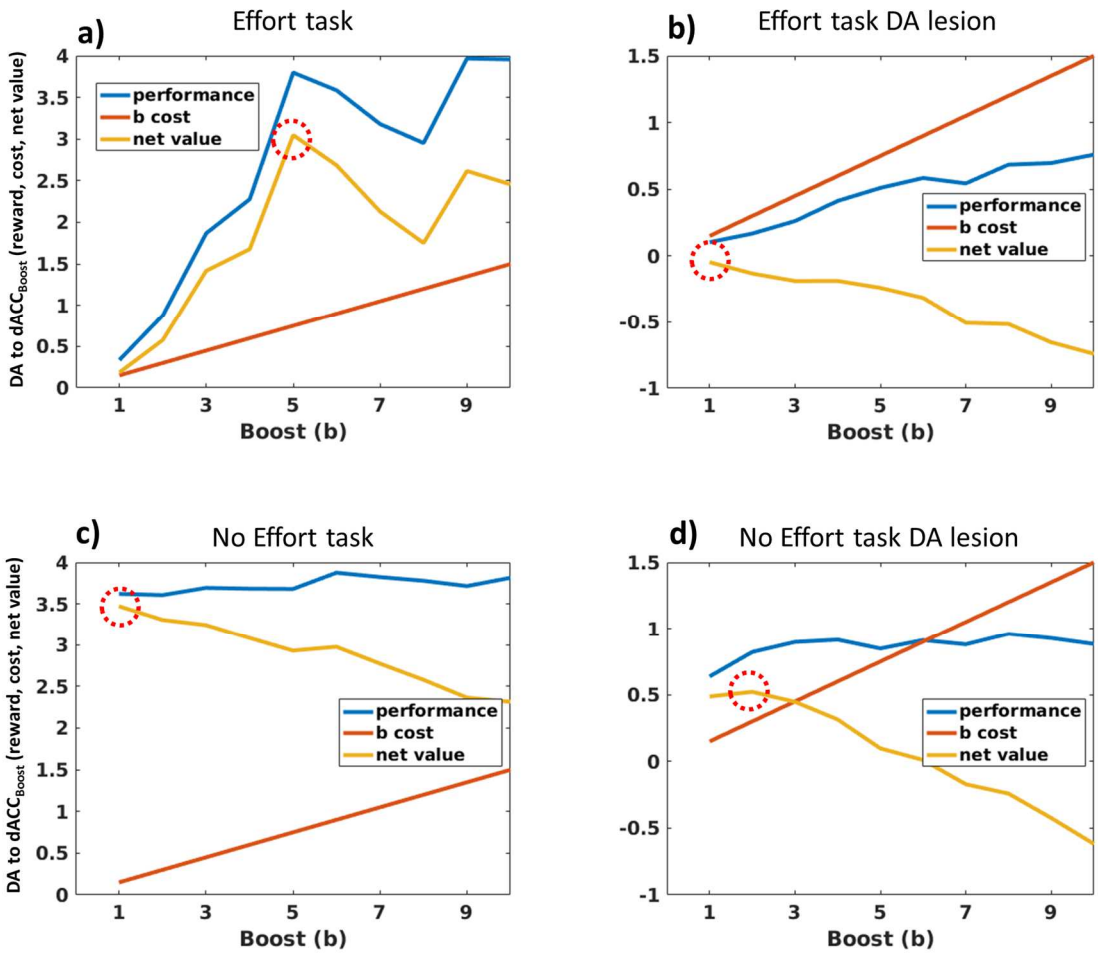


Figure 6. Cost-benefits plots and optimal control of *b* in the dACC$_{Boost}$ module. To obtain these plots we systematically clamped *b* at several values (from 1 to 10, x axis

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of each plot) and then we administered the same paradigms of Figure 5b (all the combinations Effort x DA lesion). In all the pots, y axis represents simultaneously behavioural performance in terms of average reward (for blue plots), boosting cost (red plots) and net value (performance – boost cost, in other words, Equation 7b). To minimize computation time, we use for these simulations only the discrete model. **a) Effort task, no lesion.** Plot showing RML behavioural performance as a function of *b* (blue plot), boosting cost (red plot, ωb member in Equation 7b) and net value for the $dACC_{Boost}$ module (yellow plot, resulting from Equation 7b). Red dotted circle highlight the optimal *b* value which maximizes the final net reward signal received by the $dACC_{Boost}$ module. **b) Effort task, DA lesion.** Same as a, but in this case the RML was DA lesioned. Due to lower average reward signal (blue plot), the net value plot (yellow) is much lower than in a), as the cost of boosting (red plot) did not change. Red dotted circle highlights the optimal *b* value, which is lower than in a). It must be considered that, despite the optimal *b* value is 1, the average *b* (as shown in figures 5b and s11b) is biased toward higher values, as it is selected by a stochastic process (Equation 4) and values lower than 1 are not possible (asymmetric distribution). **c) No effort task, no lesion.** In this case, being the task easy, the RML reaches a maximal performance without high values of *b*, therefore the optimal *b* value is low also in this case. **d) No effort task, DA lesion.** As shown also in Figure 5b, in this case the optimal *b* value (dotted circle), is higher than in c), because a certain amount of boosting is necessary to avoid apathy (moving as an intrinsic cost) and it can give access to large rewards. Apathy is determined by the RML preference for "Stay" option, which gives no reward but it has no costs, boosting can help avoiding apathy.

Simulation 2b: performance recovery after DA lesion, in cost/benefit trade off conditions

In DA lesioned subjects, the preference for HR option can be restored by removing the difference in effort between the two options (Walton et al., 2009), in other words by removing the critical trade-off between costs and benefits. In these simulations, we

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

show how the RML can recover a preference toward HR options when exposed to the same experimental paradigms used in rats.

Simulation Methods

The same DA lesioned subjects of Simulation 2a were exposed to either a No Effort task (where both the option required low effort) or a Double Effort task (where both the options required a high effort) (Figure 7a-b). All other experimental settings were identical to those of Simulation 2a.

Simulation Results and Discussion

DA-lesioned RML performance recovers immediately when a No Effort task is administered after the Effort task (Figure 7c), in agreement with animal data by Walton et al. (2009; Figure 7d). This result shows that performance impairment after DA lesion is not due to learning deficit (although partial learning impairment must occur due DA role in learning), but rather to down-regulation of catecholamines boosting, decided by $dACC_{Boost}$. When boosting-up catecholamines, in particular Ne that discounts costs (Equation 2), is not crucial to perform optimally the task (as both the options do not require a big effort), performance immediately recovers. An easy task does not require a strong behavioural energization, therefore the information about the high reward location is sufficient to execute correctly the task.

The same performance recovery occurs also in a task where both options are effortful (Double Effort task, Figure 7e), again in agreement with experimental data (Figure 7f). Also in this case, when there is no trade-off between costs and benefits (both the options are the same in terms of effort), the information about the high reward location is sufficient to correctly execute the task, although there is a reduced catecholamine boosting. Nonetheless, differently from the previous scenario, here emerges a new phenomenon: *apathy* (percentage of "Stay", Figure 7e-f). Indeed, both the RML and animals often refuse to engage in the task, and rather than working hard to get the high reward (whose position is well known) they prefer to remain still. Apathic behaviour in this experiment is more evident than in Figure 5e-f, because both RML and animals are forced to make an effort to get a reward, while in Simulation 2a (Figure 5e-f) they could opt for the low effort-low reward choice.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

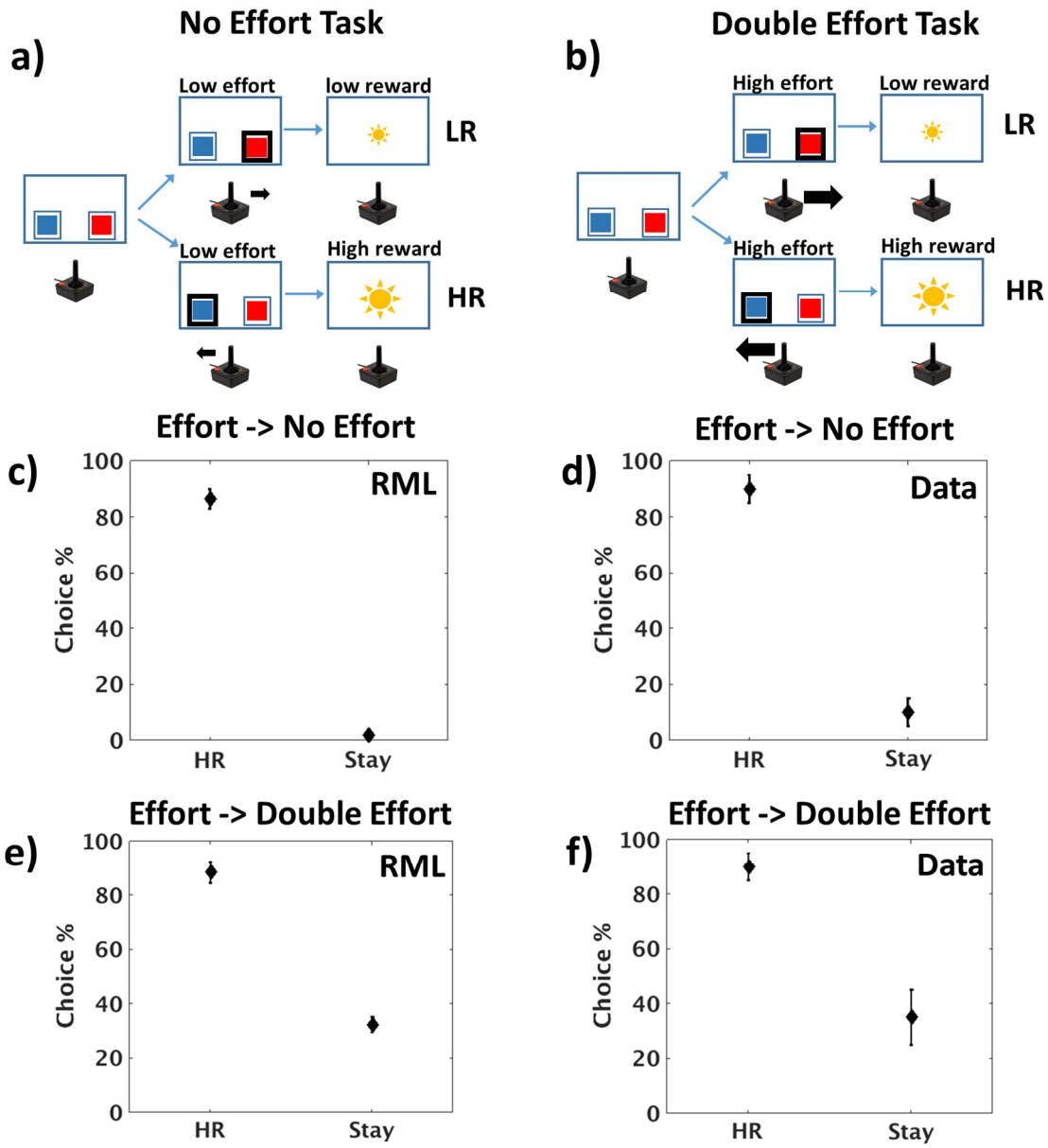


Figure 7. Recovery of HR option preference in DA lesioned subjects. **a)** No Effort task, consists in two possible choices, both requiring low effort to be executed (small black arrow), one leading to a high reward (large sun; HR), the other to a low reward (LR). **b)** Double Effort task, where both options implied high effort. **c-d)** Recovery of the preference for HR option (HR/(HR+LR)) when a No Effort task is administered after an Effort task session (Effort → No Effort), in both RML and animals (mean percentage ± s.e.m.). **e-f)** Same phenomenon when a Double Effort session follows an Effort one (Effort → Double Effort), in both RML and animals. Note that in this case

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the amount of "Stay" choices (Stay/number of trials) increased, simulating the emergence of apathic behaviour. Animal data from Walton et al., 2009.

Simulation 2c: Adapting cognitive effort 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 also 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. Given the dynamical nature of this simulation (we used a dynamical model of WM, and the task implies an intra-trial delay when information must be retained), we used in this case only the dynamical version of the RML.

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. FROST dynamics simulates the effect of memory loads on information coding,

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

with a decrement of coding precision proportional to memory load (i.e. the number of spatial locations to be maintained in memory). This feature allows to simulate the increment of behavioural errors when memory load increases (Ashby et al., 2005). The external LC output (*Ne*) improves the signal gain in the FROST DLPFC neurons, increasing the coding precision of spatial locations retained in memory (Equation s21 in Supplementary Methods), thus improving behavioural performance. 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 8a), 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 modulates *Ne* release as a function of memory load, in order to optimize performance (Figure 8b, 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 8c, 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 8c, accuracy, right panel; lesion × memory-load interaction: $F(2,22) = 8.6$, $p = 0.0017$).

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

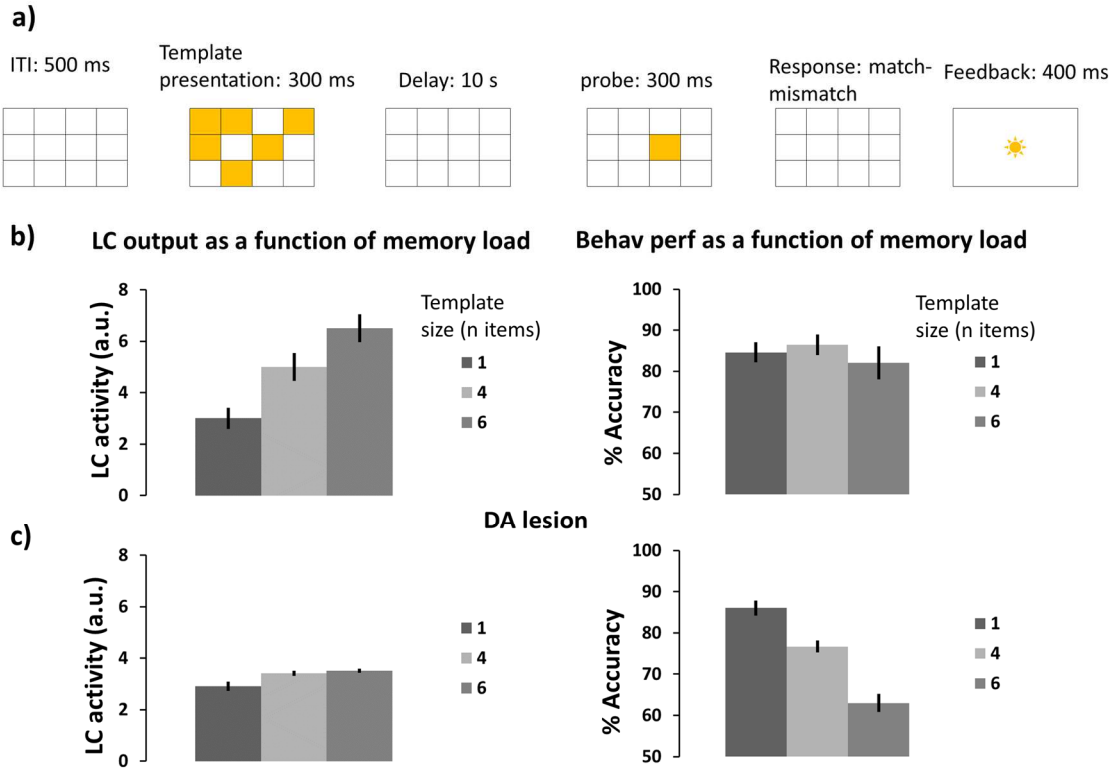


Figure 8. a) Delayed Matching-to-sample task: events occurring in one trial. **b)** Left: LC activity (*Ne* signal) as a function of memory load (number of items presented in the template). Right: behavioural performance as a function of memory load. **c)** LC activity (*Ne* signal) and behavioural performance 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 obtain adaptive

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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, literature suggests a sharp distinction between higher-order conditioning in classical versus instrumental paradigms. Indeed, although it is possible to train animals to execute complex chains of actions to obtain a reward (instrumental higher-order conditioning, Pierce and 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; Denny and Ratner, 1970). 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, no 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 (Simulation s8 in: "Supplementary Results: dynamical model"). 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: Higher-order classical conditioning

Equation 7a (and its dynamical homologous Equation s5b) expresses a progressive linking of DA response to conditioned stimuli. This is due to the max(v()) term in Equation 7a and the time derivative of *v* in Equation s5b for the dynamical version of RML. The latter equation simulates also the progressive DA shifting from reward period to cue period in classical conditioning (Schultz et al., 1993), described in Simulation 3c (Figure s8; paragraph "Supplementary Results: dynamical model"). As VTA can vigorously respond to conditioned stimuli, it is natural to wonder whether a

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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, for unknown reasons, classical higher-order conditioning is probably impossible to obtain in animal paradigms (Denny and Ratner, 1970; O'Reilly et al., 2007). We thus investigate what happens in the model in such a paradigm.

Simulation Methods

We first administered a first-order classical conditioning. 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 (R) 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. This aspect of VTA dynamics is naturally captured by the dynamical version of the RML (closer to neurophysiology than the discrete version), because at each order of conditioning, the cue-locked signal is computed as the temporal derivative of reward prediction unit activity (Equation s5b), losing part of its power at each conditioning step. In the discrete version of the model, this aspect is expressed in Equation 7a, where (like in TD-learning) the cue-locked DA response is scaled by a positive real number smaller than one (ρ). This mechanism implies a steep decay of the conditioning effectiveness of non-primary rewards, as at each order of conditioning, the reinforcing property of cues is lower and lower. From the ecological viewpoint, it makes sense that the weaker is the link between a cue and a primary reward, the weaker should be its conditioning effectiveness. Nonetheless, as we describe in the following paragraph, this phenomenon is in some way counteracted in instrumental conditioning.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

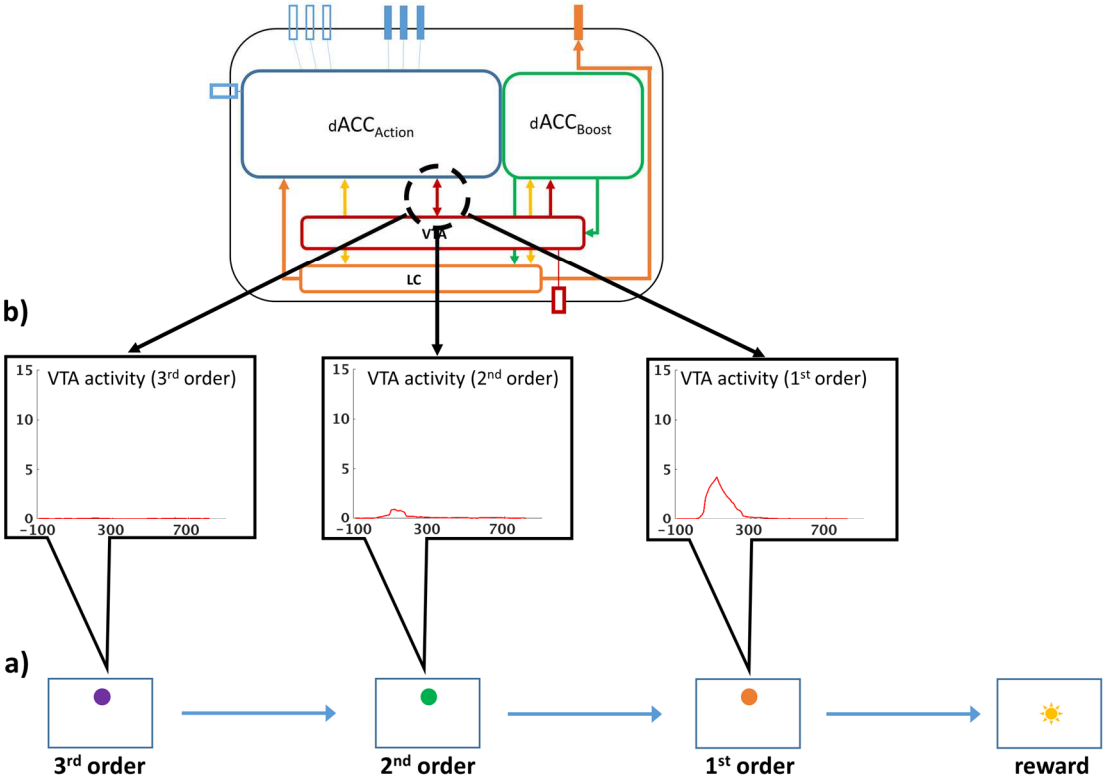


Figure 9. a) Experimental paradigm for higher-order classical conditioning. Sequence of conditioned stimuli (colored disks) followed by primary reward (sun). **b)** VTA activity locked to each of conditioned stimuli. Dashed black circle indicates where the plotted signals were recorded from (see also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

Simulation 3b: 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 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 2). The training procedure was the same as for higher-order classical conditioning. We first administered a first-order instrumental conditioning (2-armed bandit task). Then, we

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 7b) and the decision of boosting is selected only when it can result in a substantial performance improvement (in terms of achieved rewards). 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 does not 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 (Equation 7b and its dynamic homologous Equation s18), and the $dACC_{Boost}$

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

module decides not to boost, with a low VTA activation for conditioned stimuli. This explains the strong limitations in establishing higher-order classical conditioning, and shows how decisions about effort exertion are involved in higher-order conditioning.

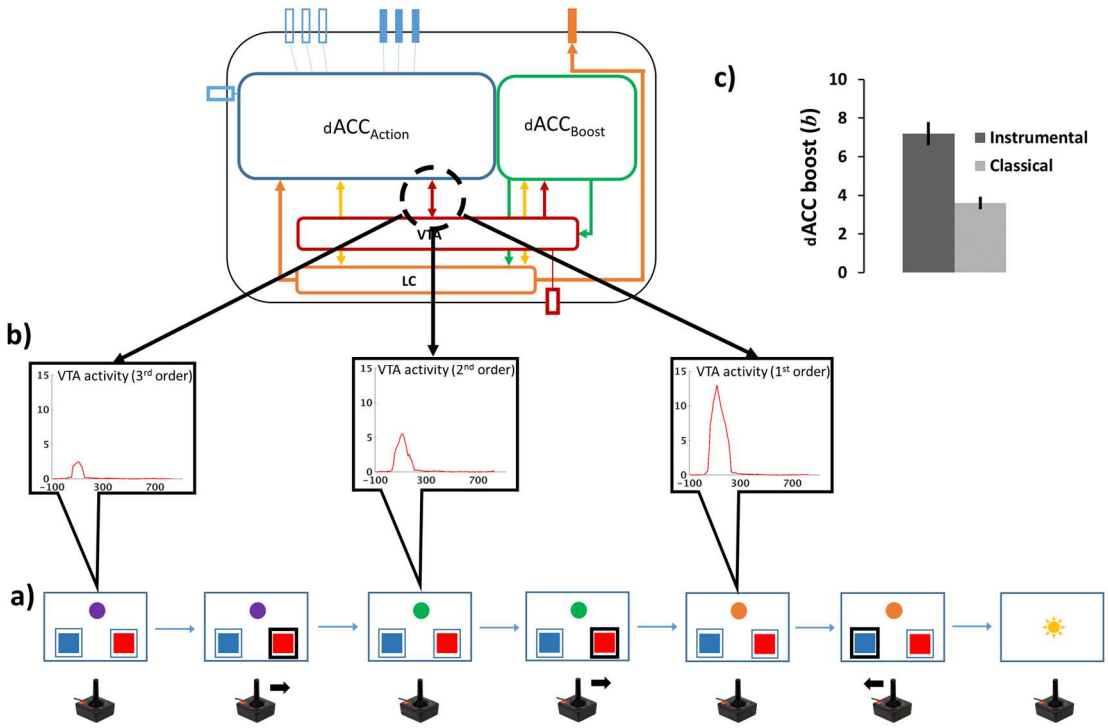


Figure 10. VTA dynamics during higher order instrumental conditioning. **a)** Events occurring during a sequence of correct choices in the task represented also in Figure 2. See Supplementary Methods for details. **b)** Cue-locked (colored disk indicating the environment state) VTA activity. Dashed black 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).

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

General Discussion

We proposed a novel account of the role of dACC in cognition, suggesting that its elusive computational role can be understood by broadening the theoretical view to a recurrent macro-circuit including the 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 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). The RML provides an explicit theory and neuro-computational architecture of how autonomous control of learning rate can emerge from dACC-LC interaction. We propose that the dACC provides raw data (in the form of RL signals) about environment statistical structure to the LC, which estimates optimal learning rate by using Bayesian inference. This explains why both structures are necessary for optimal control of flexibility (Behrens et al., 2007; Nassar et al., 2012), and why the dACC activity seems to be related to RL computation (Silvetti et al., 2013b) while the LC activity to volatility estimation (Silvetti et al., 2013a).

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-c). 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 (both primary and non-primary), introducing for the first time the dimension of subjectivity in RL. The reward signal is not anymore an "objective"

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

feedback from the environment, but it can be proactively modulated to improve behavioural performance by increasing the value of actions implying effort (Equation 7a, simulations 2a-c) or the value of non-primary rewards (simulations 3a-b). The latter mechanism, in particular, allowed to explain why higher-order conditioning is possible in instrumental but not in classical paradigms.

In the RML these three meta-learning processes are related and integrated (Figure 1c). For example, dynamic control of learning rate (λ) is based on RL signals from dACC modules (v and δ). Learning rate modulation influences both decision-making for action selection (a) and for boosting control (b). The latter modulates in parallel both LC and VTA, changing both effort investment (Ne) and reward signals (DA). Catecholamines modulation changes behavioural performance, influencing agent-environment interaction and then influencing back LC control over learning rate.

The model also contains two main limitations. First, DA plays a role only in learning. Experimental results suggest that DA is involved also in performance directly (e.g. attention and WM) (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.

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 the RML the main idea

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of the dACC as a state-action-outcome predictor. The RVPM is a subset of the dynamical version of the RML model (Dynamical Model Description, in Supplementary Methods). This implies that the RML can simulate also the results obtained by the RVPM (e.g., conflict monitoring, error likelihood estimation), extending even further the amount of empirical data that can be explained by this framework. Although the RML goes beyond these earlier works, by implementing meta-learning and higher-order conditioning, it shares with them the hypothesis that PE plays a core role for learning and decision-making. Indeed, we hypothesize that PE is a ubiquitous computational brick, which allows both dACC operations (equations 1 and 3) and the approximation of optimal learning rate in the LC (equations 6a-d).

Recent computational neuroscience of RL and decision-making focused on hierarchical architectures. For instance, Alexander and Brown (2015) proposed a hierarchical RL model (based on their previous PRO model), where hierarchical design is implemented within the dACC, unfolding 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; 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.

In line with the recent interest for hierarchical RL, also Holroyd and McClure (2015) proposed a model exploiting hierarchical RL architecture (the HRL), 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. This model can explain a wide variety of data about task selection and decision-making in cognitive and physical effort regulation.

The RML differs from these two models for the following reasons. i) Although higher-order conditioning in the RML is based, to some extent, on a hierarchical organization of reward signals (where DA signals linked to lower-level conditioned stimuli work as reward proxies to train higher-level stimuli), the RML lacks a genuine hierarchical structure. Its dynamics is emergent from the circular

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

interaction between cortical and subcortical circuits, allowing meta-learning. ii) The RML provides mechanistic explanations to experimental findings from a broader range of domains (from effort modulation to higher-order conditioning), including findings that were earlier explained by the predecessor model RVPM (e.g. error processing, error likelihood estimation, etc.). iii) Both HER and HRL needed ad hoc parameter tuning to simulate different experimental results, while the RML (in both its dynamical and discrete versions) required just one single parameter set.

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. A second model by Verguts (2017) described how dACC could implement cognitive control by functionally binding two or more brain areas by bursts of theta waves, whose amplitude would be proportional to the level of control. This theory describes how but not when (and neither how much) control should be exerted. The mechanisms proposed in the RML are an excellent complement to this theory, hypothesizing how and to what extent the dACC itself can decide to modulate theta bursts amplitude.

Finally, Khamassi et al. (2011) also hypothesizes a role of dACC in meta-learning. The authors proposed a neural model (embodied in a humanoid robotic platform) where the temperature of the action selection process (i.e. the parameter controlling the trade-off between exploration and exploitation) was dynamically regulated as a function of PE signals. Like in the RML, dACC plays both a role in reward-based decision-making and in autonomous control of parameters involved in decision-making itself. Differently from the RML, this model provided a more classical view on PE origin, which were generated by the VTA and not by the dACC like in the RML. Moreover, the mechanism proposed for temperature control was reactive to overall environmental variance (PE), lacking the capability to disentangle noise from volatility.

Concerning control of learning rate, 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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

are computationally hard (in particular for optimal Bayesian solutions, e.g. Behrens et al. 2007), and the need of hyper-parameters specifying a priori the statistical structure of the environment (Kalman, 1960; Mathys et al., 2011). The latter feature reflects the peculiarity of these models of needing an ad-hoc hierarchical forward model representing the environment. At the top of this hierarchy, the experimenter defines a-priori crucial characteristics about the environment (like the precision of the probability function describing environmental volatility). At the best of our knowledge, the only Bayesian model able to estimate volatility without the need of specifying hyper-parameters is the one by Behrens et al. (2007), which works only for binary outcomes. Also Wilson et al., (2013) proposed an approximate Bayesian estimator that is based on PE, like the RML, without the need of specifying a forward model of the environment. However, the authors provided a solution for one subclass of volatility estimation problems (the change-point problems) and also in this case, an a priori hyper-parameter describing volatility (the process variance) was needed. In contrast, the RML provides an explicit neurophysiological theory on how near-optimal control emerges from the dialogue between dACC and LC, without the need of an a priori forward model and of information about environmental volatility itself. Indeed, our hyper-parameter α represents the minimal assumption that noise variance occurs at higher frequencies than process variance (volatility). This derives from the fact that the RML is not built hierarchically, but instead different interlocking modules influence one another, creating a circular interaction. Moreover, the RML can adapt learning rate in any kind of problem (binary, continuous, etc.), and finally, it integrates approximate Bayesian optimization with other cognitive functions, like effort control and higher-order conditioning.

Experimental predictions

The flexibility of RML, and the explicit neurophysiological hypotheses on which it is based, allow several experimental predictions. Here we list some potential experiments deriving from RML predictions. The first three could potentially falsify the model.

First, the RML architecture suggests that PE signal are generated by the dACC and then they converge toward the brainstem nuclei. This hypothesis implies two experimental predictions following dACC lesion: a) disruption of DA dynamics in

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

higher-order conditioning, coupled with impossibility of higher-order instrumental training; b) disruption of LC dynamics related to learning rate control, coupled with behavioural flexibility impairment.

Second, a neurophysiological prediction concerns 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 3a-b).

Third, in Simulation 2a (Figure 5b), the DA-lesioned RML predicts stronger dACC activation during an easy task (without effort) in presence of a high reward. This mechanism has been interpreted as a compensatory phenomenon allowing to avoid apathy (i.e. refusal to engage the task) if a small effort can make available a big reward. This is an explicit experimental prediction that could be tested both in animal paradigms and in DA impaired humans.

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 2c) 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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

effort is required by the environment; Figure 3b). 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.

Future perspectives

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).

Finally, the neural units composing our model are designed as stochastic leaky integrators, allowing the RML to function in continuous time and in the presence of noise. These features are crucial to make a model survive outside the simplified

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

environment of trial-level simulations, and make possible to simulate behaviour in the real world, like, for example, in robotics applications (in preparation). RML embodiment into robotic platforms could be useful for both neuroscience and robotics. Indeed, testing our model outside the simplified environment of computer simulations could reveal model weaknesses that are otherwise hidden. Moreover, closing the loop between decision-making, body and environment (Pezzulo et al., 2011) is important to have a complete theory on the biological and computational basis of decision-making in the mammalian brain. At the same time, the RML could suggest new perspectives on natural-like flexibility in machine learning, helping, for example, in optimizing plasticity as a function of environmental changes.

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.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

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.

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

Denny M, Ratner S (1970) Comparative psychology: Research in animal behavior. Oxford: Dorsey Press.

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.

Frank MJ, Seeberger LC, O'Reilly R C (2004) By carrot or by stick: cognitive

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

reinforcement learning in parkinsonism. Science (80-) 306:1940–1943.

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.

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, Lallée S, Enel P, Procyk E, Dominey PF (2011) Robot cognitive control with a neurophysiologically inspired reinforcement learning model. Front Neurorobot 5:1.

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–

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

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

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.

Pezzulo G, Barsalou L, Cangelosi A (2011) The mechanics of embodiment: A dialog on embodiment and computational modeling. Embodied and.

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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

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.

Shenhav A, Straccia MA, Cohen JD, Botvinick MM (2014) Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. Nat Neurosci 17:1249–1254.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 (2017) Binding by Random Bursts: A Computational Model of Cognitive Control. J Cogn Neurosci 29:1103–1118.

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.

Wilson RC, Nassar MR, Gold JI (2013) A Mixture of Delta-Rules Approximation to

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Bayesian Inference in Change-Point Problems Behrens T, ed. PLoS Comput Biol 9:e1003150.

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.

bioRxiv preprint doi: https://doi.org/10.1101/130195; this version posted August 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

- Bayesian flexibility of control

Lesione LC distrugge control LR