Meta-Reinforcement Learning reconciles surprise, value and control in the anterior cingulate cortex.

Tim Vriens[1], Eliana Vassena[2,3], Giovanni Pezzulo[1], Gianluca Baldassarre[1] & Massimo Silvetti[1*]

[1] Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy

[2] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands

[*]Corresponding author: massimo.silvetti@istc.cnr.it

Abstract

The role of the dorsal anterior cingulate cortex (dACC) in cognition is a frequently studied yet highly debated topic in neuroscience. Most authors agree that the dACC is involved in either cognitive control (e.g. voluntary inhibition of automatic responses) or monitoring (e.g. comparing expectations with outcomes, detecting errors, tracking surprise). A consensus on which theoretical perspective best explains dACC contribution to behaviour is still lacking. In a recent neuroimaging study, the experimental predictions of two prominent models formalizing the cognitive control hypothesis (Expected Value of Control, EVC) and the monitoring hypothesis (Predicted Response Outcome, PRO) have been tested using a behavioural task involving both monitoring and cognitive control mechanisms. The results indicated that of the two tested models, only the PRO model effectively predicted the dACC activity, indicating surprise tracking for performance monitoring as the key sole underlying mechanism, even when cognitive control was required by the task at hand. These findings challenged the long-standing and established cognitive control hypothesis of dACC function and opened a theory crisis: the proposed surprise-monitoring hypothesis indeed cannot account for a wide array of previous experimental findings evidencing dACC activation in tasks requiring cognitive control *without* involving monitoring or surprise. Here we propose a novel hypothesis on dACC function that integrates both the monitoring and the cognitive control perspective in a unifying coherent framework, based on meta-Reinforcement Learning. Our model, the Reinforcement Meta Learner (RML), optimizes cognitive control - as in control models like EVC- by meta-learning based on tracking surprise - as in monitoring models like PRO. We tested RML experimental predictions with the same behavioural task

used to compare the PRO and EVC models, and showed that RML predictions on dACC activity matched PRO predictions and outperformed EVC predictions. However, crucially, the RML simultaneously accounts for both cognitive control and monitoring functions, resolving the theoretical impasse about dACC function within an integrative framework. In sum, our results suggest that dACC function can be framed as a meta-learning optimiser of cognitive control, providing an integrative perspective on its roles in cognitive control, surprise tracking, and performance monitoring.

Introduction

Humans constantly face complex decisions, ranging from selecting one out of more available options (like choosing between an apple or a cupcake) to choosing the amount of cognitive and bodily resources we want to invest to achieve a goal (effort allocation). Formally, these decisional processes are aimed at solving a trade-off between minimizing the cost of investing cognitive (or physical) resources and maximizing the gain from investing a certain amount of resources[1–5]. In order to optimize this trade-off, it is sometimes necessary to expend some (cognitive) effort early in order to receive a larger reward later. In experimental psychology, the mental processes deputed to control the level of cognitive effort we deploy to achieve a goal are referred to as *cognitive control.* This process allows us to select more effortful options when we anticipate higher rewards from them. A typical example of cognitive control is the ability to inhibit habitual responses when these are not appropriate for the current situation (e.g. changing the route to your workplace when there are roadworks). The dorsal anterior cingulate cortex (dACC) and the surrounding cortical areas within the medial prefrontal cortex (MPFC) are known to be involved in decision-making and cognitive control processes (e.g. [6]), but the underlying computational mechanisms are still unclear, above all considering that their activation has been routinely found in many different experimental paradigms investigating several aspects of higher-order cognition[7–18]. Ironically, this area has been labelled as the Rorschach test for neuroscientists[19].

The most recent and effective attempts to find a unified theory of dACC function have led to two competing frameworks: the performance monitoring framework[7,20] and the Expected Value of Control framework[21] (EVC). The performance monitoring framework is based on Reinforcement Learning (RL) and proposes that the dACC function is mainly aimed at performance monitoring by computing prediction error signals by comparing the expected and the actual outcome of actions and/or environmental states. Prediction error signals are essential to update expectations about future action-outcomes associations. This theory was implemented in two –independently developed - computational models, the Predicted Response Outcome model (PRO) by Alexander & Brown[7] and the Reward Value and Prediction Model (RVPM) by Silvetti et al.[20] These models showed that the combination of action-outcome expectations and the relative prediction error signals was sufficient to

explain many of the experimental findings about dACC function, from error detection to monitoring conflict between competing responses[22].

The EVC[21] states that the dACC is involved in estimating the value of exerting *cognitive control* during a specific task, and in selecting the optimal control signal, i.e. the control signal that maximises the estimated value. This framework provided a theoretical explanation for another portion of the literature that extensively showed dACC activity during anticipation of cognitively effortful tasks and the trade-off between the advantages of exerting cognitive control and its intrinsic cost [17,23–25].

Both the monitoring and cognitive control frameworks bear significant merits in offering mechanistic and integrative explanations of dACC functions. These frameworks have been directly compared in a recent fMRI study[26]. The task used by the authors – the speeded decision-making task – consisted of a two-armed bandit where the value difference between the options was variable (Figure 1a), so that the difficulty of selecting the optimal option varied from easy (when the value difference was large) to hard (when the value difference was small). Participants were asked to select the best option under time pressure. The authors demonstrated that this speeded decision-making task engaged both monitoring and cognitive control functions, making it well suited to compare predictions from both EVC and PRO models about the dACC activity during the task. The empirical results confirmed that the dACC activity was best simulated by the PRO model, while it could not be captured by the EVC. As the PRO model solely uses performance monitoring to model dACC activity, the authors concluded that dACC function can be parsimoniously explained by performance monitoring mechanisms (involving prediction and prediction error), without postulating additional cognitive control optimization mechanisms, even in tasks requiring cognitive control[27]. These results therefore question the active role of the dACC in both cognitive and physical effort optimization, which are part of cognitive control.

Unfortunately, these results lead to an undesirable situation where no comprehensive theory on the dACC function survives the whole range of experimental testing. Indeed, a growing body of literature reports activity in the dACC linked to cognitive control (or more generally during anticipation of cognitive effort) in the absence of surprise or prediction errors [3,24,28] or while controlling for all possible sources of surprise[29,30]. This means that the PRO model, which successfully explained the dACC activation in the speeded decision-making task, cannot account for the entire body of cognitive control and cognitive effort literature showing the involvement of the dACC. Conversely, the EVC model could explain many cognitive control effects in the dACC, but failed to predict the fMRI results from the speeded decision-making task. In sum, neither model could provide an integrative account of dACC function, and this seems to be a theoretical impasse unsolvable by further comparison of the experimental predictions of PRO and EVC models.

Here, we show that an alternative theoretical framework overcomes this impasse by accounting for the roles of dACC in both cognitive control and performance monitoring: The Reinforcement Meta-Learner model[31,32]. The RML belongs to a class of Reinforcement Learning algorithms (named meta Reinforcement Learning (meta-RL)) able to adapt their internal parameters as a function of environmental challenges[33,34]. The RML combines some of the features of PRO and EVC models within the perspective of meta-RL, and in previous studies, it already provided a computational account of dACC function from both cognitive control and performance monitoring perspectives[31,32,35–37]. In this paper, we compare the RML predictions on the speeded decision-making task with those from PRO and EVC models and we show that the RML matches the PRO model in simulating the dACC activity - without ad hoc parameters tuning or increasing model complexity - and outperforms the EVC. These findings resolve the theoretical impasse generated by the competition between the two previous theories, explaining the central role of the dACC in both monitoring and cognitive control with an integrative perspective.

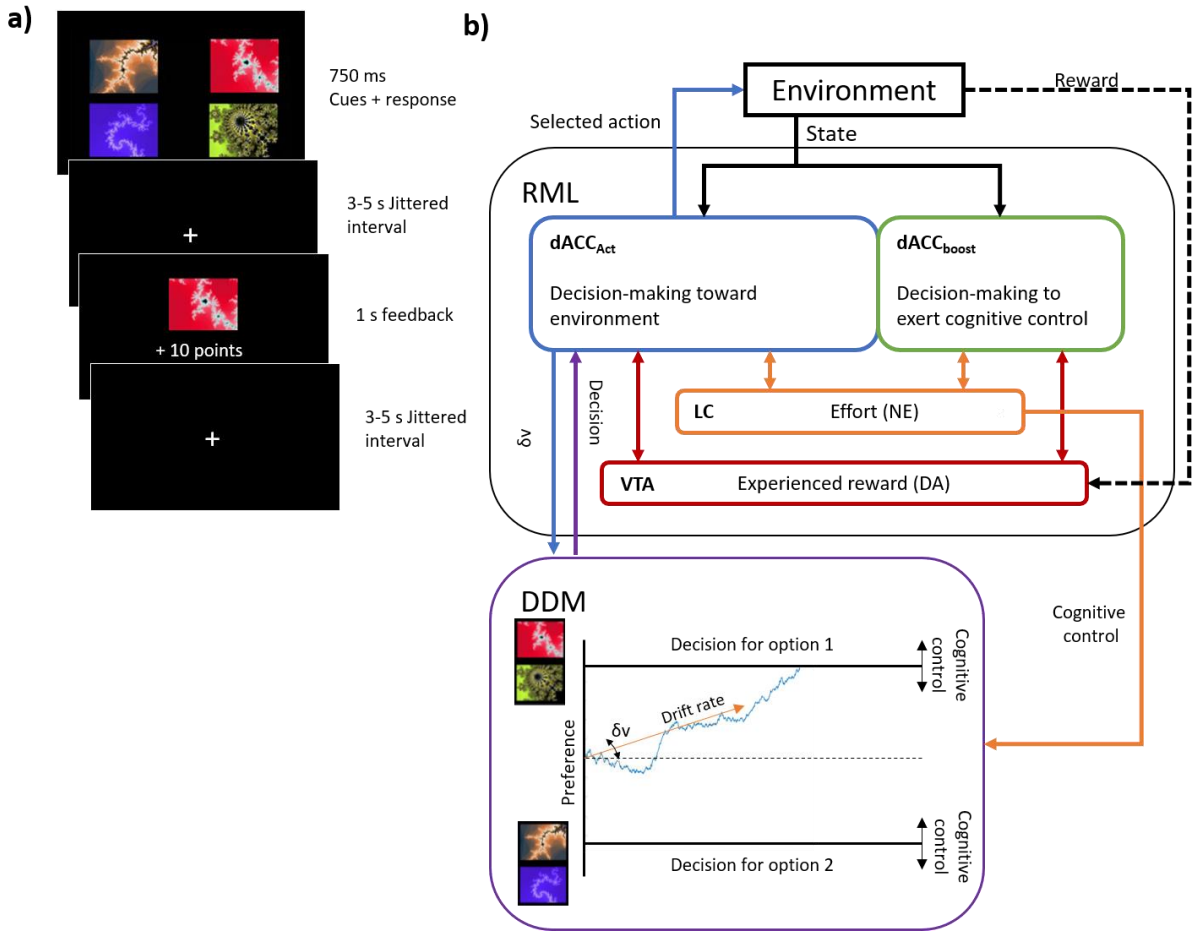

Figure 1: Task setup and RML representation: a*: The setup of the task performed by the participants in Vassena et al.[26]: participants select one of two sets of two images They receive rewards corresponding to one of the*

images selected from the chosen set at random. ***b****: Schematic representation of the model: RML receives input from the environment (about rewards and environmental states), and controls a task-specific module (the DDM), which helps in task execution.* δv: difference in the expected value of the two options, whose absolute value determines the drift rate, while its sign determines the drift direction (up or down).

Results

The Reinforcement Meta Learner (RML) model

The RML is an autonomous agent that learns and makes decisions (related to both motor behaviour and cognitive functions) to maximize net value, i.e. reward discounted by the cost of motor and cognitive effort (see Supplementary Information for a detailed description). The model has been validated across multiple tasks and research domains[31,32,35,36,38], and was used in this work without parameter tuning (keeping all the parameters as in the original papers describing the model). From the neurophysiological perspective, the RML simulates a cortical-subcortical macrocircuit including the dACC and two brainstem neuromodulatory nuclei: the ventral tegmental area (VTA) and the locus coeruleus (LC) (Figure 1b). The RML architecture assumes that the dACC activity during decision-making can be explained by three different components: the value of a decision (motor or cognitive), the surprise generated by the environmental feedback following this decision, and the level of cognitive control selected by the agent. Decision-making strategies relative to action selection and cognitive control levels (policies) are learned by agent-environment interaction, and based on approximate Bayesian learning implemented as Kalman filtering[39]. These three components of dACC activity are processed by two separate modules (Figure 1b), an $dACC_{action}$ *module* – dedicated to motor action selection – and a $dACC_{boost}$ *module* – dedicated to control optimization.

The $dACC_{boost}$ module performs control over motor and cognitive functions, by upregulating or downregulating activity in VTA and LC. When faced with a decision, the RML first decides on what amount of control to exert during the decision (termed '*boost'*). When the optimal boost level is selected, the $dACC_{boost}$ module sends a signal to the VTA and LC, which in turn influence both the dACC modules. As the optimal policy for selecting the boost signal is learned, and the boost signal itself influences the learning process, we define the search for an optimal boost policy as meta-learning. Boosting the LC module promotes effortful motor actions and enhances information processing in other brain structures (cognitive effort) while boosting the VTA module improves reward signals and learning from non-primary rewards. The boost, however, implies an intrinsic cost, and the boost module implements meta-learning by dynamically optimizing the trade-off between performance improvement and boosting cost. In this work, we will equate the boost signal to cognitive control. This

definition implies some simplification, as the boost signal can also influence the motivational component of decision-making, via VTA modulation, aside from cognitive and physical effort via LC modulation.

The $dACC_{action}$ module evaluates the options the agent has in the current environment based on expected effort and reward and selects the optimal action directed toward the environment. Action selection in the $dACC_{action}$ module is influenced by the LC input, which promotes effortful actions, and by the VTA input, which influences the expected value of actions. In this way, the action selection and value learning of the $dACC_{action}$ is indirectly modulated by the policy learned by the $dACC_{boost}$ module (meta-learning for optimal action selection).

The RML works as a task-independent optimizer of motor and cognitive decision-making. This is possible because the RML can be connected to external, task-specific, modules (e.g. a neural network simulating visuospatial processing). The activity of the external module is controlled via LC signals, and its output is directed back toward the $dACC_{action}$ module to provide the RML with the perceptual and/or motor features needed for a specific task. As the task used by Vassena et al. uses a speeded decision-making task, and the RML is unable to simulate reaction times by itself, we connected the RML to an external module simulating decisions by accumulating evidence over time, therefore enabling the model to simulate reaction times. In order to show the generalizability of our results, we replicated our simulations with two different well-known models as external modules: the drift diffusion model (DDM) as proposed by Ratcliff[40] and the dual attractor network (DAN) model by Usher & McClelland[41]. This ensured that the results could be attributed to the RML, and not to the specific external module used. In the main text, we describe the results of using the DDM as external module. These were fully replicated with the DAN module, as described in detail in the supplementary material.

The integration of the DDM and the RML (Figure 1b) is inspired by Vassena et al.[26], where the authors fitted the drift rate and the distance of decision boundaries of a DDM to their behavioural data to demonstrate that the speeded decision-making task relied on participants' cognitive control. In each trial, the DDM received two different signals from the RML as an input. The first was the difference in the expected value of the options (δv in Figure 1b) from the $dACC_{action}$ module. The value of each possible option was learned by the RML during a training session preceding the task. The absolute value of the δv signal determined the DDM drift rate, while the sign of the δv signal determined the drift direction toward one of the decision boundaries (representing the two available options, left vs right option) for that trial. In this way, a larger δv absolute value resulted in higher accuracy and faster decision[26]. The second input from the RML to the DDM consisted of the LC signal, which modulated the distance of the decision boundaries. The higher the LC activity, the closer the boundaries, and the

faster the decision. As the LC output is controlled by the boost signal from the $dACC_{boost}$ (cognitive control), the latter can therefore modulate the decision speed as a function of trial type. Once the DDM reached one decision boundary, this information was passed to the $dACC_{action}$ module that implemented the option selection; finally, both the dACC modules updated the expected value for the selected option based on the environmental feedback (reward).

In order to model the time pressure component of the task, and incentivize the RML to respond as fast as possible, we implemented a linear devaluation of the reward as a function of reaction time (supplementary Equation S13). This solution captures the fact that participants are instructed to respond as fast and as accurately as possible[26].

Simulation results

In Figure 2a we report the dACC activity measured with fMRI during the speeded decision-making task, as reported in Vassena et al.[26]. This can be described as a W-shaped function of the difference in value between the two options (quartic function with a positive leading coefficient). During the execution of the same task, the RML reproduces the W-shaped activity pattern exhibited by the dACC (Figure 2b). Using the Akaike information criterion (AIC)[42], we confirmed that the simulated dACC activity follows a quartic pattern rather than a quadratic pattern with a positive leading coefficient (Aikake weight for a quartic function > 0.999). The RML simulates the W-shaped dACC activity as the sum of two different neural signals evoked by the cue onset: the expected net value (expected reward discounted by expected costs, Figure 2c) and the cognitive control signal (boost signal in the RML, Figure 2d).

The expected net value component is dependent on the difference in cue value and follows a U-shaped function. When the values of both options are similar (around 0 on the x-axis in Figure 2c), the mean expected value of that trial type is minimal, while it is maximal for large value differences. Using the AIC[42], we found that value is better represented by a quartic function compared to a quadratic function (Aikake weight for a quartic function > 0.999). Two mechanisms cause the shape of this function. First, similar option values lead to smaller drift rates in the DDM, and then to longer RTs (see also Suppl. Fig. 2 and Equations S10, S13). Since long RTs cause reward devaluation, the RML learns that the execution of this type of trial has a lower average value (minimum in the U-shaped pattern in Figure 2c). Conversely, when the difference between option values is large, RTs are shorter on average (higher drift rates), leading to a higher expected value (maxima in Figure 2c). The second mechanism depends on the intrinsic cost of boosting. The boost function mirrors the net value function (inverted U shape, see below), and the boost cost function – which is a linear function of boost - follows the same shape (Equation S2 in Suppl. material). For this reason, the expected value is maximally

discounted in the centre and minimally discounted in the tails, contributing to a U-shaped expected value. Figure 2d shows the contribution to the RML simulation of dACC activity given by the cognitive control signal (boost), which follows an inverted U-shaped function. As the boosting level maximizes reward while minimizing the cost intrinsic to boosting, the RML increases the boost only if the consequent reward gain is worth the boost cost. When the difference between the options is small, the DDM drift rate is small. In that case, a larger boost allows for a significant improvement in performance by shrinking the decision boundaries and therefore shortening the RTs (Figure 1b, see also equation S11 in the supplementary materials). Conversely, when the task is easy, and the value difference is large, the DDM drift rate is high. In that case, RTs will be fast anyway, and increasing the boost will only marginally improve them. This leads to a reduced boost signal (cognitive control) when the value difference is large (minima in Figure 2d). Using the AIC[42], we found that boost activity is better represented by a quartic function compared to a quadratic function (Aikake weight for a quartic function > 0.999).

As shown in Figure 2 e-f, the W-shaped activity predicted by RML is in agreement with the one by PRO model, and in disagreement with the one by the EVC model, which failed to capture the human dACC activity shown in Figure 2a. Moreover, the RML provided a further experimental prediction. Indeed, the value function in Figure 2c simulated correctly the activity of another region of the MPFC besides the dACC: The vMPFC (red cluster in Figure 2h), which is well known to be associated with value estimation[43,44].

Besides the difference in simulated dACC activity during the decision-making phase of the task, Vassena et al.[26] described a second dissociation between the EVC and the PRO model, about the predicted dACC activity evoked by the absolute difference between the overall reward available in a trial and the long-term average reward (reward-driven surprise). We simulated this activity using the RML and found that both RML and PRO models correctly predicted the dACC activity - thanks to their capability to code for surprise – while the EVC failed in this prediction because it lacks a surprise-coding mechanism (Figure 2g). Vassena et al.[26] show that this function is situated in the mid-cingulate cortex, by analysing a surprise contrast in the fMRI data (green cluster in Figure 2h).

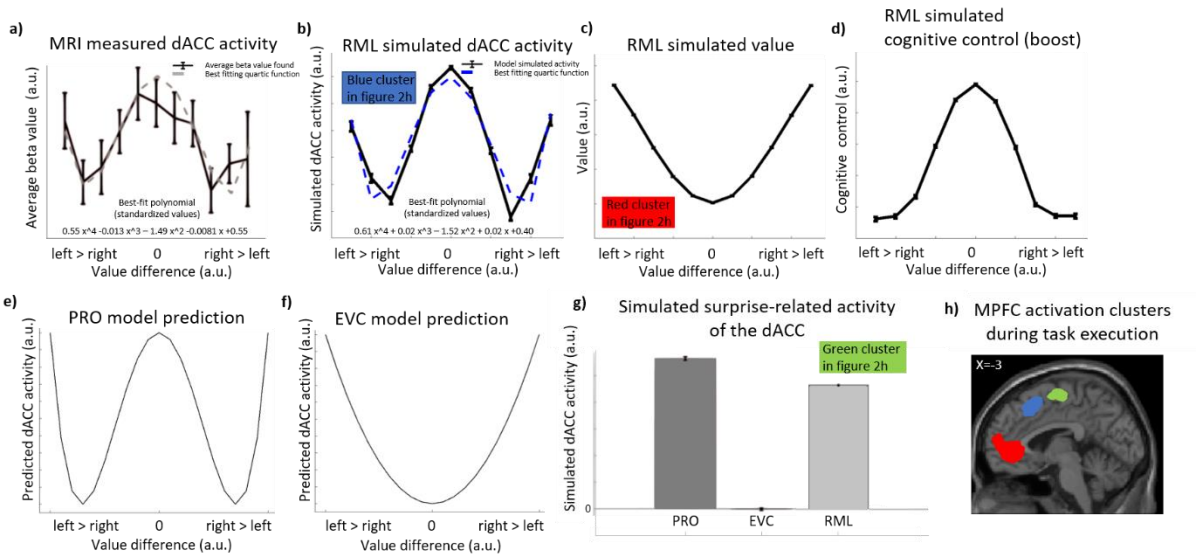

Figure 2: RML results. a*: The MRI results from Vassena et al.*[26] *(adapted). The dACC activity is shown in the black line, while the grey, dashed line shows the best fitting quartic function to this data* ***b****: The dACC activity as simulated by the RML (black line), and the best fitting quartic function (blue, dashed line). This activity is the sum of the value (panel c) and the cognitive control (panel d).* ***c****: The value component of the RML-simulated activity.* ***d:*** *the cognitive control simulated by the RML model.* ***e****: The PRO model prediction as mentioned in Vassena et al.* [26]*: quartic function with a positive leading coefficient.* ***f****: The EVC model prediction as stated by Vassena et al.*[26]*: quadratic function with a positive leading coefficient.* ***g****: Model predictions of the dACC activity relative to the absolute difference of the received reward from the average reward (average surprise). The results for the PRO and EVC model are from Vassena et al.* [26]*. Note that the dACC activity simulated by the RML model is plotted on a different scale than the dACC activity simulated by the PRO model.* ***h:*** *Activation clusters within the MPFC (modified from Vassena et al.* [26] *). Blue: dACC activation as a mixture of value and cognitive control, RML prediction in panel b; Red: vMPFC value-based activation, RML prediction in panel c; Green: mid-cingulate activation relative to average surprise, RML prediction in panel g.*

Discussion

In this study, we proposed the RML model as a solution to the theoretical impasse generated by the direct comparison between the PRO and EVC models about the role of the dACC in cognitive control processing[26]. This impasse raised from the PRO model effectiveness in predicting the dACC activity in a cognitive control task involving unexpected events—the speeded decision-making task utilized in Vassena et al.[26],—alongside its inability to simulate dACC activity in cognitive control tasks designed to exclude unexpected events (e.g. [3,24,28–30]).

Here we have shown that the RML matched the PRO model and outperformed the EVC model in predicting the dACC activation during the speeded decision-making task (Figure 2b,e). Moreover,

again in accordance with the PRO model but not with the EVC model, the RML predicted the surprise-driven activity in the mid-cingulate cortex (Figure 2g). Finally, the RML, but not the PRO model, correctly predicted also the vMPFC activity related to value estimation (Figure 2c), found by Vassena et al.[26] as a replication of experimental results from previous literature (e.g. [43,44]).

While both the RML and PRO models generate comparable predictions regarding dACC activity, they provide distinct underlying computational explanations. As we elucidate further below, this disparity is fundamental in resolving the theoretical impasse resulting from the work by Vassena and colleagues[26]. The PRO model generates the W-shaped dACC activity pattern as the combination of two different types of surprise signals. One surprise signal is triggered by the cue onset, and it depends on the uncertainty relative to the cue type. The other is triggered by the response onset and depends on the uncertainty related to the probability of selecting one of the two responses over the other. The PRO model therefore suggests that the dACC plays no role in the optimization of cognitive control per se but it is exclusively involved in surprise tracking. As argued above, this view is challenged by empirical evidence showing dACC activation corresponding to cognitive control intensity in tasks deliberately designed to exclude surprise[3,24,28–30]. In contrast, the RML explains the dACC activity as the sum of two different functions: The expected value of the selected action (U-shaped function in Figure 2c) and the cognitive control intensity (inverted U-shaped function in Figure 2d), suggesting a more complex function of the dACC, where surprise-based monitoring is functional to cognitive control optimization.

For this reason, the RML, differently from the PRO model, is able to explain the dACC activity also in tasks where surprise is not involved (e.g. ref. [31,32,35]). These findings solidify the RML as a computational account unifying the monitoring and control perspective during decision-making, thereby reconciling surprise, value, and cognitive control under the framework of meta-learning processes[33,34].

Another important innovation of the RML is its bio-inspired system-level perspective, where MPFC function is studied in conjunction with brainstem and midbrain catecholaminergic nuclei. This allows for modeling of the cortico-subcortical reciprocal influence and how the MPFC can orchestrate cortical and striatal functions by controlling the release of neuromodulators, broadening the scope of application of the RML framework. For example, in a previous work, the RML showed to be capable of providing a computational account of how the MPFC modulates parietal neurons - via LC activation - during a visual attention task[32]. The same mechanism can also explain the activity in dorsolateral regions of the frontal and parietal lobe that Vassena et al. described in their study[26]. Indeed, these areas are involved in the execution of the speeded decision-making task (spatial and motor control components), and the RML suggests that the MPFC optimizes their performance by neuronal gain

modulation via neuromodulatory input. Moreover, this system-level perspective allows us to formulate predictions about different regions within the MPFC, like the activity pattern of the vMPFC (Figure 2e) or of the mid-coingulate cortex (Figure 2g).

A further relevant element introduced by the RML (as compared to both PRO and EVC) is that most of the behavioural and neural predictions made by the RML emerge from the interaction of the model with the environment (*situatedness* as defined by Wilson[45] and Nolfi[46]). Furthermore, the goal of the optimization process of the RML is to find the set of actions (policy) that maximizes the net value, regardless of whether these actions are directed toward the external environment (motor) or the internal environment (cognitive control). Learning optimal actions (motor or cognitive) requires a loop between the RML and the environment, where the RML plays a *proactive* role in exploring the world. Recently, an improvement to the RML was proposed, the RML-C[32], where the situated and proactive aspects of RML are more pronounced, as the model's goal is not limited to net value maximization, but includes also the maximization of information relative to the environment. To this aim, the RML-C actively explores the environment to improve its predictions, investing motor and cognitive effort for gathering information (*intrinsic motivation*[47,48]), beyond the utilitarian perspective of net value maximization. The situated and proactive perspective that the RML proposes about the dACC role – and more in general of the MPFC – seems to be soundly grounded also on the neurophysiology of this area, which evidences large neural populations coding for both motor and visceromotor functions (for a review: [22]). Differently from the RML, the PRO model embraces a pure monitoring perspective, where the dACC plays a passive role of "Critic", mainly dedicated to the monitoring of action-outcome contingencies. Similarly, the EVC model - as implemented by Vassena et al.[26] - selects the control intensity as a function of the option values and the control cost, working as a decision-making algorithm where information about the environment and the action-outcome contingencies is fully known. However, a more recent implementation of the EVC, the Learned Value of Control (LVOC) model[49], has been designed for learning the optimal policies of control through model-environment interactions in stationary conditions. Yet, contrary to the RML and PRO models, the LVOC does not include explicit monitoring of surprise, making the LVOC model unable to simulate the dACC data from Vassena et al.[26]

In summary, in this work, we have shown that the meta-learning perspective represents a general solution for the understanding of the MPFC function and that it is capable of explaining empirical data from a larger domain set if compared with the PRO and EVC models alone. Future research in this domain should investigate also the comparison of the above frameworks with other, alternative perspectives on the computational basis of cognitive control, based on active inference[50,51],

such as the proposal that cognitive control results from the monitoring of the deviation from prior beliefs about cognitive actions so that cognitive effort is exerted to override habits[52].

Methods

A deatiled description of simulation methods and data analysis is reported in the Supplementary Information.

Acknowledgements

Tim Vriens is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, hosted by Università Campus Bio-Medico di Roma.

References

1. Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. Decision making and the avoidance of cognitive demand. *J. Exp. Psychol. Gen.* **139**, 665–682 (2010).
2. Salamone, J. D., Correa, M., Mingote, S. M. & Weber, S. M. Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. *Curr. Opin. Pharmacol.* **5**, 34–41 (2005).
3. Vassena, E. *et al.* Overlapping Neural Systems Represent Cognitive Effort and Reward Anticipation. *PLOS ONE* **9**, e91008 (2014).
4. Vassena, E., Deraeve, J. & Alexander, W. H. Task-specific prioritization of reward and effort information: Novel insights from behavior and computational modeling. *Cogn. Affect. Behav. Neurosci.* **19**, 619–636 (2019).
5. Walton, M. E., Rudebeck, P. H., Bannerman, D. M. & Rushworth, M. F. S. Calculating the cost of acting in frontal cortex. *Ann. N. Y. Acad. Sci.* **1104**, 340–356 (2007).
6. Heilbronner, S. R. & Hayden, B. Y. Dorsal Anterior Cingulate Cortex: A Bottom-Up View. *Annu. Rev. Neurosci.* **39**, 149–170 (2016).
7. Alexander, W. H. & Brown, J. W. Medial prefrontal cortex as an action-outcome predictor. *Nat. Neurosci.* **14**, 1338–1344 (2011).
8. Badre, D. & Wagner, A. D. Selection, integration, and conflict monitoring; assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron* **41**, 473–487 (2004).
9. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
10. Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. Conflict monitoring and cognitive control. *Psychol. Rev.* **108**, 624–652 (2001).
11. Botvinick, M. M. Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cogn. Affect. Behav. Neurosci.* **7**, 356–366 (2007).
12. Kolling, N. *et al.* Value, search, persistence and model updating in anterior cingulate cortex. *Nat. Neurosci.* **19**, 1280–1285 (2016).
13. Paus, T., Koski, L., Caramanos, Z. & Westbury, C. Regional differences in the effects of task difficulty and motor output on blood flow response in the human anterior cingulate cortex: a review of 107 PET activation studies. *Neuroreport* **9**, R37-47 (1998).

14. Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E. & Behrens, T. E. Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron* **70**, 1054–1069 (2011).

15. Shackman, A. J. *et al.* The Integration of Negative Affect, Pain, and Cognitive Control in the Cingulate Cortex. *Nat. Rev. Neurosci.* **12**, 154–167 (2011).

16. Shenhav, A., Straccia, M. A., Cohen, J. D. & Botvinick, M. M. Anterior Cingulate Engagement in a Foraging Context Reflects Choice Difficulty, Not Foraging Value. *Nat. Neurosci.* **17**, 1249–1254 (2014).

17. Vassena, E., Krebs, R. M., Silvetti, M., Fias, W. & Verguts, T. Dissociating contributions of ACC and vmPFC in reward prediction, outcome, and choice. *Neuropsychologia* **59**, 112–123 (2014).

18. Vassena, E., Holroyd, C. B. & Alexander, W. H. Computational Models of Anterior Cingulate Cortex: At the Crossroads between Prediction and Effort. *Front. Neurosci.* **11**, (2017).

19. Ebitz, R. B. & Hayden, B. Y. Dorsal anterior cingulate: a Rorschach test for cognitive neuroscience. *Nat. Neurosci.* **19**, 1278–1279 (2016).

20. Silvetti, M., Seurinck, R. & Verguts, T. Value and Prediction Error in Medial Frontal Cortex: Integrating the Single-Unit and Systems Levels of Analysis. *Front. Hum. Neurosci.* **5**, 75 (2011).

21. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron* **79**, 217–240 (2013).

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

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

24. Chong, T. T.-J. *et al.* Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLOS Biol.* **15**, e1002598 (2017).

25. Silvestrini, N., Musslick, S., Berry, A. S. & Vassena, E. An integrative effort: Bridging motivational intensity theory and recent neurocomputational and neuronal models of effort and control allocation. *Psychol. Rev.* **130**, 1081–1103 (2023).

26. Vassena, E., Deraeve, J. & Alexander, W. H. Surprise, value and control in anterior cingulate cortex during speeded decision-making. *Nat. Hum. Behav.* **4**, 412–422 (2020).

27. Vassena, E., Deraeve, J. & Alexander, W. H. Predicting Motivation: Computational Models of PFC Can Explain Neural Coding of Motivation and Effort-based Decision-making in Health and Disease. *J. Cogn. Neurosci.* **29**, 1633–1645 (2017).

28. Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E. J. & Rushworth, M. F. S. Effort-Based Cost–Benefit Valuation and the Human Brain. *J. Neurosci.* **29**, 4531–4541 (2009).

29. Engström, M., Landtblom, A.-M. & Karlsson, T. Brain and effort: brain activation and effort-related working memory in healthy participants and patients with working memory deficits. *Front. Hum. Neurosci.* **7**, (2013).

30. Engström, M., Karlsson, T., Landtblom, A.-M. & Craig, A. D. (Bud). Evidence of Conjoint Activation of the Anterior Insular and Cingulate Cortices during Effortful Tasks. *Front. Hum. Neurosci.* **8**, 1071 (2015).

31. Silvetti, M., Vassena, E., Abrahamse, E. & Verguts, T. Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLOS Comput. Biol.* **14**, e1006370 (2018).

32. Silvetti, M., Lasaponara, S., Daddaoua, N., Horan, M. & Gottlieb, J. A Reinforcement Meta-Learning framework of executive function and information demand. *Neural Netw.* **157**, 103–113 (2023).

33. Doya, K. Metalearning and neuromodulation. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **15**, 495–506 (2002).

34. Khamassi, M., Enel, P., Dominey, P. F. & Procyk, E. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. in *Progress in Brain Research* vol. 202 441–464 (Elsevier, 2013).

35. Caligiore, D., Silvetti, M., D'Amelio, M., Puglisi-Allegra, S. & Baldassarre, G. Computational Modeling of Catecholamines Dysfunction in Alzheimer's Disease at Pre-Plaque Stage. *J. Alzheimers Dis.* **77**, 275–290 (2020).

36. Goris, J. *et al.* Autistic traits are related to worse performance in a volatile reward learning task despite adaptive learning rates. *Autism* **25**, 440–451 (2021).

37. Silvetti, M., Baldassarre, G. & Caligiore, D. A Computational Hypothesis on How Serotonin Regulates Catecholamines in the Pathogenesis of Depressive Apathy. in *Multiscale Models of Brain Disorders* vol. 13 127–134 (Springer, Cham, 2019).

38. Lasaponara, S. *et al.* Temperament and probabilistic predictive coding in visual-spatial attention. *Cortex* **171**, 60–74 (2024).

39. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *ASME Ser. J. Basic Eng.* **82**, 35–45 (1960).

40. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).

41. Usher, M. & McClelland, J. L. The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.* **108**, 550–592 (2001).

42. Wagenmakers, E.-J. & Farrell, S. AIC model selection using Akaike weights. *Psychon. Bull. Rev.* **11**, 192–196 (2004).

43. Blair, K. *et al.* Choosing the Lesser of Two Evils, the Better of Two Goods: Specifying the Roles of Ventromedial Prefrontal Cortex and Dorsal Anterior Cingulate in Object Choice. *J. Neurosci.* **26**, 11379–11386 (2006).

44. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* **76**, 412–427 (2013).

45. Wilson, M. Six views of embodied cognition. *Psychon. Bull. Rev.* **9**, 625–636 (2002).

46. Nolfi, S. *Behavioral and Cognitive Robotics: An Adaptive Perspective*. (2021).

47. Baldassarre, G. & Mirolli, M. *Intrinsically Motivated Learning in Natural and Artificial Systems*. (Springer, 2013).

48. Baldassarre, G. *et al.* Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Front. Psychol.* **5**, 985 (2014).

49. Lieder, F., Shenhav, A., Musslick, S. & Griffiths, T. L. Rational metareasoning and the plasticity of cognitive control. *PLOS Comput. Biol.* **14**, e1006043 (2018).

50. Friston, K. *et al.* Active inference and epistemic value. *Cogn. Neurosci.* **6**, 187–214 (2015).

51. Pezzulo, G., Rigoli, F. & Friston, K. Active Inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol.* **134**, 17–35 (2015).

52. Parr, T., Holmes, E., Friston, K. J. & Pezzulo, G. Cognitive effort and active inference. *Neuropsychologia* **184**, (2023).