Title: Predicting motivation: computational models of PFC can explain neural coding of motivation and effort-based decision-making in health and disease

Abbreviated title: Computational models of PFC and effort-based behavior

Authors: Eliana Vassena[12], James Deraeve[2], William H. Alexander[2]

1. Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen
2. Department of experimental psychology, Ghent University

Corresponding author: Eliana Vassena, Donders Institute for Brain Cognition and Behavior, Radboud University Nijmegen, Kapittelweg 29, 6525 EN, Nijmegen, The Netherlands. +31 (0)24 3610618

email: e.vassena@donders.ru.nl

Abstract

Human behavior is strongly driven by the pursuit of rewards. In daily life, however, benefits mostly come at a cost, often requiring that effort be exerted in order to obtain potential benefits. Medial prefrontal cortex (MPFC) and dorsolateral prefrontal cortex (DLPFC) are frequently implicated in the expectation of effortful control, showing increased activity as a function of predicted task difficulty. Such activity partially overlaps with expectation of reward, and has been observed both during decision-making and during task preparation. Recently, novel computational frameworks have been developed to explain activity in these regions during cognitive control, based on the principle of prediction and prediction error (PRO model, Alexander and Brown, 2011, HER Model, Alexander and Brown, 2015). Despite the broad explanatory power of these models, it is not clear whether they can also accommodate effects related to the expectation of effort observed in MPFC and DLPFC. Here, we propose a translation of these computational frameworks to the domain of effort-based behavior. First, we discuss how the PRO model, based on prediction error, can explain effort-related activity in MPFC, by reframing effort-based behavior in a predictive context. We propose that MPFC activity reflects monitoring of motivationally relevant variables (such as effort and reward), by coding expectations, and discrepancies from such expectations. Moreover, we derive behavioral and neural model-based predictions for healthy controls and clinical populations with impairments of motivation. Second, we illustrate the possible translation to effort-based behavior of the HER model, an extended version of PRO model based on hierarchical error prediction, developed to explain MPFC-DLPFC interactions. We derive behavioral predictions which describe how effort and reward information is coded in PFC, and how changing the configuration of such environmental information might affect decision-making and task-performance involving motivation.

Introduction

Attaining a goal often requires commitment, implementation of a precise course of actions, and deployment of sufficient resources to reach successful completion. How humans fulfill this process is largely investigated in the field of cognitive neuroscience under the term of goal-directed behavior. A crucial underlying cognitive mechanism is prediction: the ability to evaluate the environment, formulate expectations about future events based on previous experiences, and finally compare such expectations with subsequent outcomes in order to update one's knowledge about the current state of the world. Furthermore, adaptive interaction with the environment entails predicting the impact of one's action and evaluating outcomes. The human Prefrontal Cortex (PFC) is the neural machinery that supports crucial mechanisms involved in goal-directed behavior (Miller & Cohen, 2001). In particular, the medial portion of PFC (MPFC, including dorsal Anterior Cingulate Cortex, dACC) is implicated in prediction and outcome evaluation (Alexander & Brown, 2011; Jahn, Nee, Alexander, & Brown, 2014), including situations in which the outcome carries value for the agent, such as a reward (Rushworth, Walton, Kennerley, & Bannerman, 2004; Silvetti, Seurinck, & Verguts, 2011, 2013; Vassena, Krebs, Silvetti, Fias, & Verguts, 2014). Notably, a variety of other functions have been attributed to MPFC (Vassena, Holroyd, & Alexander, 2017), including error monitoring, (Holroyd, Nieuwenhuis, Mars, & Coles, 2004; van Veen, Holroyd, Cohen, Stenger, & Carter, 2004), conflict detection (Botvinick, Braver, Barch, Carter, & Cohen, 2001), pain and affect processing (Bush, Luu, & Posner, 2000; Nee, Kastner, & Brown, 2011), and value-based decision-making (Rangel & Hare, 2010; Rushworth, Kolling, Sallet, & Mars, 2012; Rushworth & Behrens, 2008). Recent computational work has explained this array of empirical findings under the unifying principle of prediction and prediction error. In this framework, MPFC formulates predictions tracking stimuli, actions and outcomes, and computes a signal termed prediction error, which scales with the discrepancy between predicted and actual outcomes

(Predicted Response Outcome model, PRO model, Alexander & Brown, 2011). This mechanism allows rapid prediction updating according to environmental feedback, be it an error, a painful stimulus, or a reward. An extended version of the same model, the Hierarchical Error Representation model (HER, Alexander & Brown 2015), expands the same computational principle in a hierarchical architecture, capturing more complex high-level cognitive processes involving the interaction of MPFC and dorsolateral PFC (DLPFC), typically associated with higher-level cognitive functions such as working memory and goal-maintenance (Miller & Cohen, 2001).

The goal of this manuscript is to explore the power of these computational accounts, in terms of generating novel neural and behavioral predictions for untested contexts and populations. These frameworks have proven useful across several fields of cognition, yet they have not been put to test in the field of effortful behavior and motivation. Goal-directed behavior generally involves competing factors, including the value of the prospective goals, how much effort one is willing to exert to attain the desired goal, and preparation for the necessary effortful performance (Botvinick & Braver, 2015; Westbrook & Braver, 2013, 2015). First, we will describe MPFC involvement in effort-based behavior. Then, we illustrate how the PRO model can be generalized to the domain of motivation. We propose that MPFC activity reflects monitoring of motivationally relevant variables such as reward and required effort, instead of coding an explicit cost-benefit or choice signal per se. We illustrate novel model-based simulations, as well as theoretical predictions, which can be used to guide further empirical enquiry. We discuss how the PRO framework makes neural and behavioral predictions for clinical conditions in which motivation is impaired, such as depression and other psychiatric disorders (Treadway, Bossaller, Shelton, & Zald, 2012). Subsequently, we discuss the future directions in translating the HER model to the domain of motivation, extrapolating behavioral predictions.

From such predictions, we derive implications for measuring and potentially training motivation-related cognitive mechanisms in clinical populations.

Effort-based decision-making and performance in MPFC

Experimental manipulation of effort in behavioral and neuroimaging experiments has yielded a wealth of findings in the past decade. Typically, effort is perceived as aversive (Kool, McGuire, Rosen, & Botvinick, 2010), yet humans decide to engage in it when doing so leads to a benefit, such as a reward. In the framework of decision-making and neuroeconomics, the net value of a potential reward is discounted (i.e. decreased) by the amount of effort required to obtain the reward (Apps, Grima, Manohar, & Husain, 2015; Hartmann, Hager, Tobler, & Kaiser, 2013; Nishiyama, 2014). This seems to hold across different types of effort (Nishiyama, 2016), and guides decisions to engage one's resources in the task at hand (Kurzban, Duckworth, Kable, & Myers, 2013). Several studies in animals described the neural mechanisms underlying this cost-benefit evaluation, showing a pivotal role of MPFC in interaction with striatal and sub-cortical nuclei (Hosokawa, Kennerley, Sloan, & Wallis, 2013; Kennerley, Dahmubed, Lara, & Wallis, 2009; Rushworth & Behrens, 2008; Walton, Kennerley, Bannerman, Phillips, & Rushworth, 2006; Walton et al., 2009; Walton, Bannerman, Alterescu, & Rushworth, 2003;. Walton, Bannerman, & Rushworth, 2002; Walton, Rudebeck, Bannerman, & Rushworth, 2007). In the last few years a similar network has been characterized in humans. Lesions of MPFC may result in a condition known as akinetic mutism (Devinsky, Morrell, & Vogt, 1995), whereby patients show difficulties in initiating speech and movement, not due to an impairment of related systems, but rather to the inability or "lack of will" to start it. Electrical stimulation of the same region seems to induce a general feeling of being more motivated and more willing to persevere in effortful endeavors (Parvizi, Rangarajan, Shirer, Desai, & Greicius, 2013, although this single-case study presents some methodological caveats). More recently, neuroimaging studies have shown involvement of the

striatum and MPFC in effort-reward trade-off computations (Botvinick, Huffstetler, & McGuire, 2009; Croxson, Walton, O'Reilly, Behrens, & Rushworth, 2009; Engstrom, Landtblom, & Karlsson, 2013; Massar, Libedinsky, Weiyan, Huettel, & Chee, 2015; Mulert et al., 2008). Furthermore, expecting to perform a more cognitively challenging task is associated with increased activity in striatum and MPFC, overlapping with activity induced by the prospect of a higher reward (Krebs, Boehler, Roberts, Song, & Woldorff, 2012; Vassena, Silvetti, et al., 2014).

These results suggest a crucial contribution of MPFC to effort-based behavior, hypothesized to compute the willingness to engage in the task at hand, given that upon completion a reward will be received. This principle has been defined in recent theories (Holroyd & Yeung, 2012; Shenhav, Botvinick, & Cohen, 2013), and formalized in a computational model wherein MPFC calculates the value of boosting certain actions over others, accordingly guiding behavior in cognitive and physical tasks (Verguts, Vassena, & Silvetti, 2015). Such computations are thought to influence decision-making (Treadway, Buckholtz, et al., 2012), resource allocation, task preparation (Kurniawan, Guitart-Masip, Dayan, & Dolan, 2013; Vassena, Silvetti, et al., 2014), and response vigor (Kurniawan, Guitart-Masip, & Dolan, 2011), even at the lowest layers of the motor system (Vassena, Cobbaert, Andres, Fias, & Verguts, 2015).

In summary, growing evidence supports a pivotal role of MPFC in effort-based behavior. However, such empirical effects and theorizing efforts have so far failed to provide a precise computational characterization able to account for this line of evidence within other existing computational frameworks of MPFC function.

The Predicted Response Outcome (PRO) Model

According to the PRO model, MPFC implements two core mechanisms: 1) learning to predict the outcome of responses generated in response to environmental stimuli and 2) signaling

discrepancies between predictions and observations. Using these two primary signals as an index of MPFC activity, the PRO model has previously been shown to account for a range of effects observed in MPFC related to cognitive control and decision making, including effects of error, conflict, and error likelihood. Critically, the PRO model explains these effects without reference to the underlying affective import: feedback related to behavioral error is equivalent to feedback indicating correct behavior in the sense that both forms of feedback constitute an outcome that can be predicted on the basis of task-related stimuli. An open question, therefore, is how the PRO model might be extended to account for effects in which behavior is influenced not only by the likelihood of an event occurring, but also by the value of that event.

---------- Insert Figure 1 here ----------

Translating the PRO model to effort-based motivation

According to the PRO framework, MPFC activity encodes prediction error, resulting in increased activity for more unexpected (surprising) events. However, several studies investigating effort-based behavior report increased activity in the same region of MPFC when more effort needs to be invested (i.e. in presence of a more demanding task, Krebs et al., 2012; Vassena, Silvetti, et al., 2014). To reconcile this apparent inconsistency, we hypothesize that MPFC contribution to effort-based decision making parallels its role in cognitive control – MPFC predicts the amount of effort (as well as reward) associated with certain environmental cues, and the likelihood of the choice to engage or not in the required behavior. In other words, we propose that MPFC monitors effort cues and decisions, with the same mechanisms used to monitor the occurrence of any other stimulus and response outcome.

Decisions regarding whether to engage in an effortful task carry multiple consequences. First, the choice to perform an effortful task entails exerting actual effort in order to perform the task (regardless whether the task is performed successfully or not). Additionally, performing a task carries with it the possibility of success, in which case the subject receives positive feedback, often in the form of monetary reward. Alternately, the subject may fail to perform the task successfully, in which case negative feedback is provided indicating failure. In the simulations, such failure corresponds to not realizing the monetary reward, rather than losing money (although a loss condition could also be simulated as easily). In the framework of the PRO model, then, the outcomes predicted during choices regarding whether to engage in an effortful task are 1) the level of effort to be exerted and 2) the potential expected payoff. Furthermore, our implementation relies on two assumptions. First, greater effort is considered an aversive outcome, which generally tends to be avoided if possible (Kool et al., 2010). Second, as in the original model implementation (Alexander & Brown, 2011), outcomes can be more or less salient: increasing levels of reward and effort correspond to increasing salience in the model. This assumption is based on the observation that effort is frequently perceived as aversive, plausibly generating increased arousal level.

Under these assumptions, we simulated effort-based decision making with the PRO model. The parameter set used here was the same used in simulations reported in earlier work (Alexander & Brown, 2011, 2014), with no additions to the architecture of the model, and therefore not specifically tailored to the current context (the code is available at https://github.com/modelbrains/PRO_Effort). One should note that in this case the PRO model is not performing the task itself, but rather monitoring the choice of engaging in more or less effortful and rewarding trials (i.e. updating its predictions as a function of the experienced effort and reward, as if the task had been performed), as opposed to accepting a default option with a low reward value and no effort. In this formulation, MPFC activity reflects a monitoring

194 signal, tracking the (un)predictability of motivationally relevant variables, instead of explicitly
195 computing a cost-benefit trade off or driving choice. Related work (cf. Brown & Alexander,
196 this issue) suggests how signals generated by the PRO model may be deployed elsewhere in the
197 brain to drive choice behavior. Additionally, the adaptation of the PRO model to the context of
198 effort-based decision-making suggests that the role of MPFC is primarily in monitoring the
199 level of prospective reward and effort, and does not necessarily drive decisions to engage in a
200 proposed task, nor, once engaged, to maintain performance levels sufficient to realize
201 successful completion of a task. Rather, according with additional applications of the PRO
202 model in this issue (cfr. Brown & Alexander, this issue), signals generated by MPFC are
203 incorporated into decision processes occurring beyond cingulate itself. This interpretation of
204 MPFC function is at odds with other models of MPFC function (Holroyd & Yeung, 2012;
205 Shenhav et al., 2013), and provides a novel view of the role of MPFC in motivated behavior
206 that may be the target of future research.

207 For simulations of the effort-based decision-making task, the model was presented a compound
208 cue indicating the level of prospective reward (4 levels) and level of prospective effort (4
209 levels). Each reward level was modeled as a single input unit, as was each effort level, for a
210 total of 16 unique compound stimuli reflecting combinations of effort and reward information.
211 Following a decision to perform the task, the model received feedback related to the level of
212 reward received and the level of effort expended. The strength of the feedback signal for both
213 effort and reward was set to the level indicated by the corresponding model input (1 through 4)
214 multiplied by a constant (0.48 for reward, 0.55 for effort). The constant was selected by hand
215 to reproduce the qualitative pattern of behavioral effects reported in the literature (Klein-
216 Flügge, Kennerley, Saraiva, Penny, & Bestmann, 2015). Following a decision not to engage in
217 the effort task, feedback was set to 1/4 of the lowest reward level.

218 ---------- Insert Figure 2 here ----------

Figure 2 shows the results of the simulations. The model behavior replicates qualitatively effort avoidance tendencies of human participants (see Figure 2a): as the required effort (task difficulty) increases, the probability of engaging in the task decreases (i.e. the prediction that one will choose to engage). Plausibly, the prospect of a high reward changes this pattern: when a higher reward is expected, the probability of engaging in more effortful tasks decreases only slightly relative to low reward conditions. These behavioral predictions are consistent with empirical findings of previous studies (Apps et al., 2015; Klein-Flügge, Kennerley, Friston, & Bestmann, 2016). By looking at activity of the prediction units in the PRO model (Figure 2b), one can extrapolate quantitative predictions about expected activity in MPFC across different effort and reward conditions. According to the simulation, MPFC activity monotonically and linearly increases as a function of increased required effort (task difficulty) when reward prospect is high. However, when reward prospect is low, MPFC activity increases less steeply and only up to a certain degree of required effort, subsequently decreasing as the probability of engaging in trials with high-demand for low reward drops. To our knowledge, this neural prediction is yet to be tested and could be investigated by recording MPFC activity during effort-based decision-making when difficulty is manipulated parametrically.

Alternative models of effort-based behavior

Other theoretical and computational models have been developed to account for MPFC contribution to effort-based behavior (Shenhav et al., 2013; Verguts et al., 2015). These models present one major difference with respect to the PRO framework: they explicitly operationalize effort as a cost to be computed in MPFC. As a result, while these models work well in predicting effort-based decisions and task-performance, they do not provide an explicit computational characterization of how MPFC contributes to other empirical effects.

Verguts and colleagues (2015) assign MPFC a role in calculating the benefit of deploying effort in addition to signaling potential rewarding outcomes. Their adaptive effort investment model

operationalizes effort explicitly by implementing what the authors call "boosting". In this model, units representing MPFC activity compute the value of boosting, namely exerting the effort needed to energize a more difficult action (be it a physical action or a cognitive task). Boosting, as in exerting effort, entails a cost. If the value of boosting outweighs the cost, the more effortful action will be selected. This results in the following predicted pattern of activity: overall activity in MPFC should be higher for larger rewards, increase with increasing task-difficulty as long as the reward is worth the effort, and drop for tasks too difficult to be solved. To our knowledge, this prediction still requires empirical testing. In line with this model, Shenhav and colleagues proposed the "expected value of control theory" (EVC, Shenhav et al. 2013).This theoretical framework assigns MPFC the role of computing the value of exerting control, by combining "component computations" estimating costs, benefits, and consequences associated with control signals (Shenhav, Cohen, & Botvinick, 2016). Input signals to such computations may include error, conflict, difficulty and prediction errors signals, which may originate outside MPFC.

From the implementation point of view one should consider that the adaptive effort allocation and EVC frameworks rely on very different assumptions as compared to the PRO model. The first two place computation of the value of boosting (for cognitive or physical action in Verguts et al.) or exerting control (cognitive tasks in Shenhav et al.) in MPFC, while the PRO model places prediction and prediction error computations in MPFC. Moreover, whereas the PRO model postulates a shared underlying computational principle, adaptive effort allocation and EVC imply the coexistence of different computations (cost and value of boosting, and prediction error for the first, separate cost, benefit and consequences estimation for the second). However, further modeling work is required to extrapolate predictions, which may disentangle the models based on available empirical evidence.

The main advantage of the PRO model is parsimony: the same architecture explains effort-related effects as well as a wide variety of empirical effects previously measured in MPFC (ranging from prediction error, cognitive control, conflict and so forth, Alexander & Brown, 2011). This is not the case for the adaptive effort investment model, which is specifically tailored for effort-based behavior and is therefore not applicable in other contexts, at least in its current implementation.

One limitation of the PRO model is that it does not perform the task and is not responsible for action selection: MPFC units compute predictions and compare them with outcomes. This assumes that the reward and cost trade-off computations, and the choice to engage or not in the task at hand are implemented elsewhere. Candidate areas could potentially be other sub-regions of PFC, or possibly the basal ganglia and especially the striatum, shown in several studies in both humans and animals to contribute to effort-based decisions and task-preparation (Bailey, Simpson, & Balsam, 2016; Botvinick et al., 2009; Prévost, Pessiglione, Météreau, Cléry-Melin, & Dreher, 2010; Salamone, Correa, Farrar, & Mingote, 2007; Vassena et al., 2014). One shortcoming common to both PRO and EVC/adaptive effort allocation frameworks is that they are agnostic about cost computation. Effort is plausibly defined as a function of task-difficulty and higher effort equals higher cost. However the nature and source of such cost signal, is a topic of ongoing empirical and theoretical work (Holroyd, 2016; Kurzban et al., 2013).

Predictions and implications for clinical populations

Adaptive decision-making and energization of behavior poses a challenge in several daily life situations. In a number of psychiatric conditions, these mechanisms are impaired. Recent studies showed that decision-making regarding whether to undertake an effortful task is altered in depression, bipolar disorder and schizophrenia (Barch, Treadway, & Schoen, 2014; Culbreth, Westbrook, & Barch, 2016; Hershenberg et al., 2016; McCarthy, Treadway, Bennett, & Blanchard, 2016; Silvia et al., 2016; Silvia, Nusbaum, Eddington, Beaty, & Kwapil, 2014;

Treadway, 2016; Treadway, Bossaller, et al., 2012; Yang et al., 2014). Symptoms often include reduced willingness to exert effort, although data across different pathologies or effort types do not always align. For example, both schizophrenia and depressed patients show reduced allocation of physical effort for higher rewards as compared to controls, while evidence concerning cognitive effort is mixed (Barch, Pagliaccio, & Luking, 2016). The same authors suggest a different underlying deficit for the two conditions: depressed patients seem to show impaired hedonic processing, while schizophrenia patients tend to show impaired reinforcement learning and action selection. Moreover, effort-related deficits in schizophrenia point to an effort allocation deficit, rather than reduced effort expenditure per se (McCarthy et al., 2016; Treadway, Peterman, Zald, & Park, 2015), with patients performing less optimal decisions. Such a complex picture confirms alteration of effort-based decision-making in such clinical populations, and calls for more precise quantitative frameworks, able to identify the mechanisms underlying different impairments.

Here, we use the PRO model, adapted as described above for modeling effort-related dynamics in healthy subjects, to simulate the possible neuroetiology underlying clinical disorders, which could explain the behavioral symptoms measured in clinical samples. In the PRO model, outcome representation units may be modulated by salience (Alexander & Brown, 2011) suggesting that compromised function in clinical populations may be a result of altered perception of salient events (Alexander, Fukunaga, Finn, & Brown, 2015). Model simulations and theoretical predictions are described in Figure 3.

---------- Insert Figure 3 here ----------

These simulations use the basic architecture of the PRO model without modification as in simulation 1. In order to simulate altered function during effort-based decision-making, we assume that clinical disorders entail alterations in the processing of information related either to reward or effort information. One possible alteration driving impairment in decision-making

could be attributed to a *global salience change*: in some populations, the global salience of decision variables might be affected. Patients may be overly sensitive to the costs of engaging in a task (such as required effort, simulation 2, Figure 3b), or have reduced sensitivity to potential reward (simulation 3, Figure 3c). To simulate these hypotheses, we multiply the effort level from simulation by a factor of 2 (simulation 2), or the reward information by a factor of 0.5 (simulation 3) to reflect increased effort salience or decreased reward salience. The results of these simulations show that increasing the salience of effort and reducing the salience of reward have similar effects in the model: the probability of engaging in a task is decreased over all levels of reward and effort. The pattern of MPFC activity predicted by the model is also severely attenuated relative to control simulation: activity is slightly higher in the high reward as compared to low reward condition, but does not seem to track effort as it did in the control simulation.

Another possible alteration underlying the impairment in clinical populations might be a *mismatch*: predictions made by the model regarding effort and reward levels might not correspond to veridical experience. The model may overestimate the level of effort required (simulation 4) or underestimate the value of the reward on offer (simulation 5). The inability to accurately estimate required effort and potential reward, would generate a mismatch between prediction and outcome: predicted effort could be overestimated, leading to abnormal effort avoidance, while mismatches between predicted and experienced reward could lead to decreased motivation in performing the task. To simulate these hypotheses, effort-related feedback to the model was multiplied by a factor of 2 (simulation 4), while the valence information used for updating top-down control weights (Alexander & Brown, 2011, supplementary Figure 1) remained unchanged. The net effect is that the model's prediction of effort level exceeds the effort experienced by the model following choices to engage in an effortful task. In simulation 5, reward-related feedback to the model was multiplied by 0.5

(while valence information was unchanged), with the interpretation that the level of predicted reward did not match the experienced level. Simulation results for effort mismatch (Fig. 3d) and reward mismatch (Fig. 3e) show that such mismatches in effort and reward prediction yield qualitatively different predictions regarding behavior: overestimation of effort level leads to increased discounting of low reward offers, while behavior in high-reward conditions is relatively unaffected compared to control simulations. Conversely, underestimation of reward produces a general increase in discounting - both high and low reward conditions are discounted more heavily compared to control simulations.

To our knowledge, the hypothesized mechanisms (*global salience impairment vs. predicted/actual mismatch*) cannot be disentangled on the basis of existing data. Future experimental work to test this may incorporate a model-based fMRI experiment with patients performing an effort-based decision-making task. One could simulate model-based predictors of MPFC activity for each hypothesized mechanism on the basis of subjects' actual performance. This would show which mechanisms better explain brain activity measured in MPFC (i.e. the one giving better fit between model activity and neural data). Empirical verification in clinical populations showing impairments of effort-based behavior would shed light on potential mechanisms underlying symptoms origin and provide (in)validation for the PRO as plausible neurofunctional account of MPFC contribution to motivated behavior.

Limitations and critical aspects

Translating the PRO framework to a motivational context allows explaining effort-based behavior under a working computational model of MPFC functioning without postulating a MPFC function dedicated to explicit cost computations. However, this translation leaves some critical aspects unanswered, open for future research. First, in our conceptualization we do not distinguish between different types of effort costs, such as physical vs. cognitive effort. Here we only assume higher effort to be more salient and aversive, irrespective of its specific nature.

Previous research comparing neural circuits involved in different effortful tasks (Schmidt, Lebreton, Cléry-Melin, Daunizeau, & Pessiglione, 2012) suggests that the type of effort determined the network involved in task execution, with motor regions implicated in a physical task as opposed to parietal regions implicated in a cognitive task. In both cases, the relevant network was more active in the high effort condition. Moreover, a shared motivational hub was identified in the striatum, showing increased activity irrespective of effort type. In both animal and human research the striatum has been implicated in cost-benefit trade-offs (Botvinick et al., 2009; Salamone et al., 2016; Vassena, Silvetti, et al., 2014; Westbrook, A. & Braver, 2016), and is often co-active with MPFC. An intriguing possibility is that striatal dopamine-driven trade-off computations provide MPFC with the necessary cost signal regulating subsequent behavior, irrespective of effort type. These speculations should be investigated in further research.

Second, we do not include a mechanistic explanation of the aversive nature of effort. The neural origin of this computation is still debated in the literature. It has been proposed that perception of effort cost derives from its opportunity cost (i.e. engaging resources which could be utilized differently, Kurzban et al., 2013). A recent account hypothesizes effort cost to derive from accumulation of waste product at the neural level, resulting from using up neural resources (Holroyd, 2016). The model is currently agnostic to the origin of this signal, which we consider an avenue for future modeling and experimental work.

Third, we formulated effort-based behavior as a decision-making problem, where effort and reward are considered outcomes of the decision to engage in the task at hand. However, this does not account for monitoring ongoing effort exertion. Maintaining a certain level of vigor throughout a period (e.g. holding a grip) could be seen as the result of a series of decisions to keep engaging throughout the entire period, depending on (presumably striatal) cost and reward

392 signals fed into MPFC. This intriguing idea should be addressed in future modeling and
393 experimental work.

394 Fourth, we do not simulate MPFC activity variations within a trial. Theoretically, the PRO
395 model states that MPFC continuously predicts stimulus-outcome associations (Alexander &
396 Brown, 2014). This means that at the beginning of a trial, prior to effort or reward related
397 information being presented, the model would predict average outcomes (in the context of
398 effort-based decision-making, these predictions would converge on the mean reward and effort
399 for the overall task). Following cue presentation, this prediction would be updated when
400 experiencing the actual effort, suggesting that MPFC activity should reflect the degree by
401 which task-related cues on a specific trial diverge from the average experimental value.
402 Preliminary evidence for such a computation is reported in a Transcranial Magnetic Stimulation
403 study measuring motor-evoked potentials (Vassena et al., 2015), which showed that motor
404 cortex excitability during cue presentation was related to prediction error in expected value
405 (discrepancy between average expected value and value of the actual cue on the current trial,
406 integrating a certain degree of required effort and potential reward). However, how this result
407 speaks to MPFC contribution in the process remains to be investigated. Conversely, activity
408 following the choice regarding whether to engage with an effortful task should vary inversely
409 with the tendency of the subject to engage: subjects with a lower overall tendency to engage in
410 effortful tasks should show increased MPFC activity following choices to engage, while
411 subjects with a strong tendency for engaging should show increased MPFC activity following
412 choices not to engage. These theoretical predictions require empirical testing, and possibly
413 additional modeling work to specify them quantitatively. From the methodological point view,
414 one would need to collect fMRI data at a time scale with sufficient resolution to contrast MPFC
415 activity at both cue and outcome, or to use EEG-fMRI simultaneous recordings to localize
416 MPFC electrophysiological signature.

Effort-based decision-making and performance in DLPFC

In the existing literature, the link between DLPFC and effort-based behavior is more implicit, although it clearly emerges from the number of high-level functions implicating this region. Traditionally, DLPFC is assigned a pivotal role in supporting working memory updating and maintenance (Curtis & D'Esposito, 2003; Miller & Cohen, 2001). In recent years, evidence has accumulated showing a crucial contribution of this region to executive functions, including goal-maintenance and task-set representation (Ridderinkhof, van den Wildenberg, Segalowitz, & Carter, 2004). Activity in DLPFC is associated with maintaining stimulus representations and strategies for optimal task performance. Although several frameworks have been proposed to explain DLPFC function (Badre, 2008; Koechlin, 2014), a mechanistic account of how such representations and strategies are formed and manipulated to guide goal-directed behavior is still lacking. How DLPFC interacts with MPFC prediction signals remains unclear. Several studies investigating motivation and task-preparation report co-activation of DLPFC and MPFC (Botvinick & Braver, 2015; Chong et al., 2017; Engström, Karlsson, Landtblom, & Craig, 2014; Engstrom et al., 2013; Krebs et al., 2012; Rypma, Berger, & D'Esposito, 2002; Vassena, Silvetti, et al., 2014). Across these studies, DLPFC activity increases as a function of expected effort, task-load and working memory demands. Recently, starting from the principles outlined in the PRO model, it has been proposed that the underlying computational mechanism of DLPFC might also rely on the prediction and prediction error (Alexander & Brown, 2015). These authors proposed an updated version of the PRO model, extended in a hierarchical architecture: the Hierarchical Error Representation model (HER).

---------- Insert Figure 4 here ----------

The HER model is composed of 2 or more hierarchical layers, and each layer replicates the functional form of the PRO model. The lowest layer receives input and feedback from the environment, updating predictions via prediction error, computed as the discrepancy between

predicted and actual outcome. Additionally, the error signal also provides input to the layer above, where it is treated as a feedback signal; in other words, this higher layer learns predictions of the expected error of the lower layer, compares such prediction with the actual error signal, and updates the error prediction accordingly. This simple architecture provides a mechanistic account of how MPFC and DLPFC might interact, congruent with available empirical evidence (Alexander & Brown, 2015, 2016). The prediction error signal generated in MPFC not only results in an updated error prediction at the highest layer: this prediction is also linked to the environmental stimulus (or context), which was associated with the error. This results in a representation linking the error signal to the stimulus (or context) that preceded the error. In agreement with a substantial body of evidence, this model accounts for the primary role of MPFC in performance monitoring and error detection, and for the role of DLPFC in maintaining task-set representations providing context for MPFC function.

Future directions: translating the HER model to effort-based behavior

Despite its wide explanatory power, the HER framework has to date not been translated to the domain of motivation to accommodate for the aforementioned effort and task-load effects observed in DLPFC. In the previous sections we showed the potential of the PRO model to explain effort-related effects. Fundamentally, the HER model is an extension of the PRO model, which suggests it might be well suited for a comparable translation to the effort domain.

The aim of the current section is twofold. First we propose a theoretical explanation of how DLPFC-MPFC interaction in the context of the HER model could account for motivational effects observed in both regions. Second, we derive informal behavioral predictions from the HER model in its current formulation which can be tested in both healthy and clinical populations to further challenge the validity of the model. One should note that such interpretations and prediction are highly speculative at this stage. The purpose of this section is to provide a series of directions and predictions to drive empirical investigation of DLPFC-

MPFC contribution to effort-based behavior.

The HER model is built on the principle that error signals in MPFC are equivalent to other environmental feedback signals, and are therefore subject to the same prediction and error processes. When an error signal is unexpected, the error prediction is updated. This *error history* is stored in DLPFC as error representations linked to stimuli or environmental contexts. This implies that when the same stimulus or environmental context reoccurs, the corresponding DLPFC error representation is also reactivated. We hypothesize that this representation will in turn up-regulate MPFC activity, reinstating the signal experienced at the time of error, but this time with the purpose of exerting control to prevent the error from happening again (thus leading to a better prediction, or a successful behavioral outcome). In this formulation, the translation to a motivational context becomes evident: a performance error, for example due to task difficulty, would be signaled by increased MPFC activity, tagging that particular behavioral instance as requiring extra effort. Next time the same instance reoccurs, the reactivated error representation can provide information necessary to inform top-down control and resource allocation to result in successful task performance. Noteworthy, this speculative explanation does not require an explicit operationalization of effort or other motivational factors: thus, the HER model in its current architecture could be able to account for both prediction-related as well as effort-related signals in MPFC and DLPFC. The empirical validity of this explanatory framework is to be tested in future research, which should provide neurobiological evidence for the type of MPFC-DLPFC dynamics described above.

Besides the theoretical implications for understanding PFC circuitry, the model relies on assumptions that require empirical testing. The hierarchical structure of the HER model is consistent with other accounts of PFC function, postulating the existence of a cortical rostro-caudal hierarchical gradient (Badre, 2008; Koechlin, 2016). According to these theories, caudal regions of PFC encode more concrete representations (action-related, or more recent in time),

while more rostral regions encode more abstract representations (task-sets, rules, context or information further in the past to be maintained). This is implemented in the HER model, wherein a typical simulation of a working memory task, different items to be stored in working memory are encoded at different levels of the hierarchy (depending on order of processing, see for example the 12AX task simulations in Alexander and Brown, 2015). An underlying assumption is serial processing, not only for series of stimuli, but also for complex stimuli composed by different stimulus features. When placed in the context of motivation and decision-making literature, this assumption is quite relevant: most of the experiments referenced above combine motivationally salient information of different types, such as required effort and available rewards, presenting it simultaneously. The empirical question remains open as to whether such simultaneous presentation results in simultaneous or serial processing of the presented information sources, and to date this question has not been addressed. The HER framework hypothesizes that such features would be processed serially in a specific and preferred order. Simulations showed that altering this order, by imposing a non-preferred order, can impact performance (Alexander & Brown, 2015).

Presenting motivationally salient information prior to task performance typically influences accuracy, reaction time and task preparation in several tasks requiring cognitive control (Aarts & Roelofs, 2010; Boehler, Schevernels, Hopf, Stoppel, & Krebs, 2014; Janssens, De Loof, Pourtois, & Verguts, 2016; van den Berg, Krebs, Lorist, & Woldorff, 2014; Vassena, Silvetti, et al., 2014). When applied to the domain of effort-based behavior, the order hypothesis predicts that altering order of processing of reward and effort information might result in a shift in perceived subjective value, and consequently affect (improve or deteriorate) performance.

Predictions and implications for clinical populations

These theoretical predictions naturally stem from the HER model, and empirical testing of their validity carries relevant implications. First, testing these predictions will (dis-)prove the

validity of the assumptions underlying the model. Second, if altering order of processing can alter decision-making, one could test the potential of such manipulation to improve dysfunctional decision making, for example concerning health-related behavior such as physical exercise and eating habits. Third, if altering order of processing can alter performance, one could devise optimal ways to reconfigure available motivational information to improve cognitive performance, for example in educational and school settings. Lastly, all of the above have important implications for translational research and potential applications in clinical populations affected by disorders of motivation.

To date, the predictions listed above have not been empirically tested. It is however useful to speculate on the mechanisms, which could underlie such effects. One plausible explanation involves salience. If effort and reward information is processed serially, the order of processing when presentation is simultaneous may be influenced by the respective salience of informative cues. Patients with depression typically show reduced willingness to exert effort to obtain a reward: in other words they are more effort-avoidant as compared to controls (Treadway, Buckholtz, Schwartzman, Lambert, & Zald, 2009; Yang et al., 2014). One possible reason could be the overestimation of the amount of the required effort, which would result in an unfavorable overall value, leading to the decision of not engaging in the task. Similarly, reward information could be underestimated, thus reducing the worth of the final value. Note that these hypotheses are in line with what was formulated and simulated with the PRO model in the previous section of this manuscript, where we hypothesized impairment in perceived salience of effort and reward. What is particularly relevant with respect to the order effects, is the possibility of intervention: motivational impairment could derive from altered perception of saliency; manipulating order of presentation may enforce a specific order of processing, artificially increasing or decreasing saliency of effort and/or reward information; by tuning this manipulation, one might be able to determine the optimal configuration restoring normal

perception of salience. Ideally, this process would result in increasing the willingness to exert effort in exchange for reward, thus counteracting the typical behavioral pattern of anhedonia, a core symptom in depression (Silvia et al., 2014; Treadway, Bossaller, et al., 2012). Critically, alterations in effort- and reward-based decision making have also been reported in other psychiatric conditions such as bipolar disorder and schizophrenia (Barch et al., 2014; Fervaha et al., 2013; Gold, Waltz, & Frank, 2015; Hershenberg et al., 2016) and pre-clinical traits of apathy (Bonnelle et al., 2015), although showing different patterns of impairment. Testing the predictions derived from the HER model across different clinical samples could provide insights on shared and dissociable underlying etiopathogenetic mechanisms; moreover, such deeper theoretical understanding could foster development of behavioral treatments aimed at improving decision-making and behavioral outcomes for these patients in daily life.

General discussion

This manuscript reviews the theoretical frameworks provided by the PRO and HER models, modeling the neurofunctional architecture of MPFC and DLPFC. Such models have originally been developed based on the core principles of prediction and prediction error to explain empirical effects found in these regions. Here we discussed how these models may generalize to the domain of motivation, focusing on effort-based behavior. We show that effort effects in MPFC can be successfully accounted for by the PRO model, which provides further predictions regarding behavior and neural activity in both healthy and clinical populations. Furthermore, we discuss the potential translation of the HER model to the domain of effort-based behavior, which accounts for empirical effects measured in DLPFC, and provides interesting empirical predictions regarding the effect of order of processing on decision-making and task-performance: if these predictions are borne out, such effects could lead to the development of useful interventions to influence altered perception of salience of effort and reward information in clinical population, potentially improving abnormal behavior.

One primary goal of this manuscript is to emphasize the importance of exploiting precise theoretical frameworks to derive predictions to test experimentally. The first advantage of such mathematically precise frameworks resides in the ability to explain several behavioral and neural effects observed in a brain region under the same computational principle. The second advantage is the possibility to generate new predictions based on the same model, which can translate to contexts to date untested or different populations. This feature is particularly useful to guide further theory-driven empirical inquiry. In a scientific age where empirical tools proliferate, basing experimental research on strong *a priori* hypotheses has become a necessary condition to allow drawing statistically meaningful and generalizable conclusions. Finally, such theoretical rigor and quantitative predictive precision provide a great tool to test potential translational applications, with broad explanatory power for understanding the neurobiology of disease.

Acknowledgements

WHA and JD were supported by FWO-Flanders Odysseus II Award #G.OC44.13N.

EV was supported by the Marie Sklodowska-Curie action with a standard IF-EF fellowship, within the H2020 framework (H2020-MSCA-IF2015, Grant number 705630).

References

Aarts, E., & Roelofs, A. (2010). Attentional Control in Anterior Cingulate Cortex Based on Probabilistic Cueing. *Journal of Cognitive Neuroscience*, *23*(3), 716–727. https://doi.org/10.1162/jocn.2010.21435

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. https://doi.org/10.1038/nn.2921

590 Alexander, W. H., & Brown, J. W. (2014). A general role for medial prefrontal cortex in event
591 prediction. *Frontiers in Computational Neuroscience*, *8*, 69.
592 https://doi.org/10.3389/fncom.2014.00069
593 Alexander, W. H., & Brown, J. W. (2015). Hierarchical Error Representation: A Computational
594 Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex. *Neural Computation*,
595 *27*(11), 2354–2410. https://doi.org/10.1162/NECO_a_00779
596 Alexander, W. H., & Brown, J. W. (2016). Frontal cortex function derives from hierarchical
597 predictive coding. *bioRxiv*, 76505. https://doi.org/10.1101/076505
598 Alexander, W. H., Fukunaga, R., Finn, P., & Brown, J. W. (2015). Reward salience and risk
599 aversion underlie differential ACC activity in substance dependence. *NeuroImage.*
600 *Clinical*, *8*, 59–71. https://doi.org/10.1016/j.nicl.2015.02.025
601 Apps, M. A. J., Grima, L. L., Manohar, S., & Husain, M. (2015). The role of cognitive effort in
602 subjective reward devaluation and risky decision-making. *Scientific Reports*, *5*, 16880.
603 https://doi.org/10.1038/srep16880
604 Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the
605 frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200.
606 Bailey, M. R., Simpson, E. H., & Balsam, P. D. (2016). Neural substrates underlying effort,
607 time, and risk-based decision making in motivated behavior. *Neurobiology of Learning*
608 *and Memory*, *133*, 233–256. https://doi.org/10.1016/j.nlm.2016.07.015
609 Barch, D. M., Pagliaccio, D., & Luking, K. (2016). Mechanisms Underlying Motivational
610 Deficits in Psychopathology: Similarities and Differences in Depression and
611 Schizophrenia. *Current Topics in Behavioral Neurosciences*, *27*, 411–449.
612 https://doi.org/10.1007/7854_2015_376

613 Barch, D. M., Treadway, M. T., & Schoen, N. (2014). Effort, anhedonia, and function in
614 schizophrenia: reduced effort allocation predicts amotivation and functional impairment.
615 *Journal of Abnormal Psychology*, *123*(2), 387–397. https://doi.org/10.1037/a0036299
616 Boehler, C. N., Schevernels, H., Hopf, J.-M., Stoppel, C. M., & Krebs, R. M. (2014). Reward
617 prospect rapidly speeds up response inhibition via reactive control. *Cognitive, Affective &*
618 *Behavioral Neuroscience*, *14*(2), 593–609. https://doi.org/10.3758/s13415-014-0251-5
619 Bonnelle, V., Veromann, K.-R., Burnett Heyes, S., Lo Sterzo, E., Manohar, S., & Husain, M.
620 (2015). Characterization of reward and effort mechanisms in apathy. *Journal of*
621 *Physiology-Paris*, *109*(1–3), 16–26. https://doi.org/10.1016/j.jphysparis.2014.04.002
622 Botvinick, M., & Braver, T. (2015). Motivation and Cognitive Control: From Behavior to
623 Neural Mechanism. *Annual Review of Psychology*, *66*(1), 83–113.
624 https://doi.org/10.1146/annurev-psych-010814-015044
625 Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict
626 monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.
627 Botvinick, M. M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human
628 nucleus accumbens. *Cognitive, Affective & Behavioral Neuroscience*, *9*(1), 16–27.
629 https://doi.org/10.3758/CABN.9.1.16
630 Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior
631 cingulate cortex. *Trends in Cognitive Sciences*, *4*(6), 215–222.
632 Chong, T. T.-J., Apps, M., Giehl, K., Sillence, A., Grima, L. L., & Husain, M. (2017).
633 Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLoS*
634 *Biology*, *15*(2), e1002598. https://doi.org/10.1371/journal.pbio.1002598
635 Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E., & Rushworth, M. F. (2009).
636 Effort-based cost-benefit valuation and the human brain. *Journal of Neuroscience*,
637 *29*(14), 4531–4541. https://doi.org/10.1523/JNEUROSCI.4515-08.2009

638 Culbreth, A., Westbrook, A., & Barch, D. (2016). Negative symptoms are associated with an
639 increased subjective cost of cognitive effort. *Journal of Abnormal Psychology*, *125*(4),
640 528–536. https://doi.org/10.1037/abn0000153
641 Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during
642 working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423.
643 https://doi.org/10.1016/S1364-6613(03)00197-9
644 Devinsky, O., Morrell, M. J., & Vogt, B. A. (1995). Contributions of anterior cingulate cortex
645 to behaviour. *Brain: A Journal of Neurology*, *118 (Pt 1)*, 279–306.
646 Engström, M., Karlsson, T., Landtblom, A.-M., & Craig, A. D. B. (2014). Evidence of Conjoint
647 Activation of the Anterior Insular and Cingulate Cortices during Effortful Tasks.
648 *Frontiers in Human Neuroscience*, *8*, 1071. https://doi.org/10.3389/fnhum.2014.01071
649 Engstrom, M., Landtblom, A.-M., & Karlsson, T. (2013). Brain and effort: brain activation and
650 effort-related working memory in healthy participants and patients with working memory
651 deficits. *Frontiers in Human Neuroscience*, *7*, 140.
652 https://doi.org/10.3389/fnhum.2013.00140
653 Fervaha, G., Graff-Guerrero, A., Zakzanis, K. K., Foussias, G., Agid, O., & Remington, G.
654 (2013). Incentive motivation deficits in schizophrenia reflect effort computation
655 impairments during cost-benefit decision-making. *Journal of Psychiatric Research*,
656 *47*(11), 1590–1596. https://doi.org/10.1016/j.jpsychires.2013.08.003
657 Gold, J. M., Waltz, J. A., & Frank, M. J. (2015). Effort Cost Computation in Schizophrenia: A
658 Commentary on the Recent Literature. *Biological Psychiatry*, *78*(11), 747–753.
659 https://doi.org/10.1016/j.biopsych.2015.05.005
660 Hartmann, M. N., Hager, O. M., Tobler, P. N., & Kaiser, S. (2013). Parabolic discounting of
661 monetary rewards by physical effort. *Behavioural Processes*, *100*, 192–196.
662 https://doi.org/10.1016/j.beproc.2013.09.014

Hershenberg, R., Satterthwaite, T. D., Daldal, A., Katchmar, N., Moore, T. M., Kable, J. W., & Wolf, D. H. (2016). Diminished effort on a progressive ratio task in both unipolar and bipolar depression. *Journal of Affective Disorders*, *196*, 97–100. https://doi.org/10.1016/j.jad.2016.02.003

Holroyd, C. B. (2016). The waste disposal problem of effortful control. *Motivation and Cognitive Control*, 235–260.

Holroyd, C. B., Nieuwenhuis, S., Mars, R. B., & Coles, M. G. (2004). Anterior cingulate cortex, selection for action, and error processing. *Cognitive Neuroscience of Attention*, 219–231.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, *16*(2), 122–128. https://doi.org/10.1016/j.tics.2011.12.008

Hosokawa, T., Kennerley, S. W., Sloan, J., & Wallis, J. D. (2013). Single-neuron mechanisms underlying cost-benefit analysis in frontal cortex. *Journal of Neuroscience*, *33*(44), 17385–17397. https://doi.org/10.1523/JNEUROSCI.2221-13.2013

Jahn, A., Nee, D. E., Alexander, W. H., & Brown, J. W. (2014). Distinct regions of anterior cingulate cortex signal prediction and outcome evaluation. *NeuroImage*, *95*, 80–89. https://doi.org/10.1016/j.neuroimage.2014.03.050

Janssens, C., De Loof, E., Pourtois, G., & Verguts, T. (2016). The time course of cognitive control implementation. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-015-0992-3

Kennerley, S. W., Dahmubed, A. F., Lara, A. H., & Wallis, J. D. (2009). Neurons in the frontal lobe encode the value of multiple decision variables. *Journal of Cognitive Neuroscience*, *21*(6), 1162–1178. https://doi.org/10.1162/jocn.2009.21100

687 Klein-Flügge, M. C., Kennerley, S. W., Friston, K., & Bestmann, S. (2016). Neural Signatures
688 of Value Comparison in Human Cingulate Cortex during Decisions Requiring an Effort-
689 Reward Trade-off. *The Journal of Neuroscience*, *36*(39), 10002–10015.
690 https://doi.org/10.1523/JNEUROSCI.0292-16.2016
691 Klein-Flügge, M. C., Kennerley, S. W., Saraiva, A. C., Penny, W. D., & Bestmann, S. (2015).
692 Behavioral Modeling of Human Choices Reveals Dissociable Effects of Physical Effort
693 and Temporal Delay on Reward Devaluation. *PLOS Comput Biol*, *11*(3), e1004116.
694 https://doi.org/10.1371/journal.pcbi.1004116
695 Koechlin, E. (2014). An evolutionary computational theory of prefrontal executive function in
696 decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*,
697 *369*(1655). https://doi.org/10.1098/rstb.2013.0474
698 Koechlin, E. (2016). Prefrontal executive function and adaptive behavior in complex
699 environments. *Current Opinion in Neurobiology*, *37*, 1–6.
700 https://doi.org/10.1016/j.conb.2015.11.004
701 Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the
702 avoidance of cognitive demand. *Journal of Experimental Psychology. General*, *139*(4),
703 665–682. https://doi.org/10.1037/a0020198
704 Krebs, R. M., Boehler, C. N., Roberts, K. C., Song, A. W., & Woldorff, M. G. (2012). The
705 involvement of the dopaminergic midbrain and cortico-striatal-thalamic circuits in the
706 Integration of reward prospect and attentional task demands. *Cerebral Cortex (New York,*
707 *NY)*, *22*(3), 607–615. https://doi.org/10.1093/cercor/bhr134
708 Kurniawan, I. T., Guitart-Masip, M., Dayan, P., & Dolan, R. J. (2013). Effort and valuation in
709 the brain: the effects of anticipation and execution. *Journal of Neuroscience*, *33*(14),
710 6160–6169. https://doi.org/10.1523/JNEUROSCI.4777-12.2013

711 Kurniawan, I. T., Guitart-Masip, M., & Dolan, R. J. (2011). Dopamine and effort-based
712 decision making. *Frontiers in Neuroscience*, *5*, 81.
713 https://doi.org/10.3389/fnins.2011.00081
714 Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of
715 subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6).
716 https://doi.org/10.1017/S0140525X12003196
717 Massar, S. A. A., Libedinsky, C., Weiyan, C., Huettel, S. A., & Chee, M. W. L. (2015).
718 Separate and overlapping brain areas encode subjective value during delay and effort
719 discounting. *NeuroImage*, *120*, 104–113.
720 https://doi.org/10.1016/j.neuroimage.2015.06.080
721 McCarthy, J. M., Treadway, M. T., Bennett, M. E., & Blanchard, J. J. (2016). Inefficient effort
722 allocation and negative symptoms in individuals with schizophrenia. *Schizophrenia*
723 *Research*, *170*(2–3), 278–284. https://doi.org/10.1016/j.schres.2015.12.017
724 Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function.
725 *Annual Review of Neuroscience*, *24*(1), 167–202.
726 https://doi.org/10.1146/annurev.neuro.24.1.167
727 Mulert, C., Seifert, C., Leicht, G., Kirsch, V., Ertl, M., Karch, S., … Jäger, L. (2008). Single-
728 trial coupling of EEG and fMRI reveals the involvement of early anterior cingulate cortex
729 activation in effortful decision making. *NeuroImage*, *42*(1), 158–168.
730 https://doi.org/10.1016/j.neuroimage.2008.04.236
731 Nee, D. E., Kastner, S., & Brown, J. W. (2011). Functional heterogeneity of conflict, error,
732 task-switching, and unexpectedness effects within medial prefrontal cortex. *NeuroImage*,
733 *54*(1), 528–540. https://doi.org/10.1016/j.neuroimage.2010.08.027
734 Nishiyama, R. (2014). Response effort discounts the subjective value of rewards. *Behavioural*
735 *Processes*, *107*, 175–177. https://doi.org/10.1016/j.beproc.2014.08.002

736 Nishiyama, R. (2016). Physical, emotional, and cognitive effort discounting in gain and loss
737 situations. *Behavioural Processes*, *125*, 72–75.
738 https://doi.org/10.1016/j.beproc.2016.02.004
739 Parvizi, J., Rangarajan, V., Shirer, W. R., Desai, N., & Greicius, M. D. (2013). The will to
740 persevere induced by electrical stimulation of the human cingulate gyrus. *Neuron*, *80*(6),
741 1359–1367.
742 Prévost, C., Pessiglione, M., Météreau, E., Cléry-Melin, M.-L., & Dreher, J.-C. (2010).
743 Separate valuation subsystems for delay and effort decision costs. *Journal of*
744 *Neuroscience*, *30*(42), 14080–14090. https://doi.org/10.1523/JNEUROSCI.2752-10.2010
745 Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice.
746 *Current Opinion in Neurobiology*, *20*(2), 262–270.
747 https://doi.org/10.1016/j.conb.2010.03.001
748 Ridderinkhof, K. R., van den Wildenberg, W. P., Segalowitz, S. J., & Carter, C. S. (2004).
749 Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action
750 selection, response inhibition, performance monitoring, and reward-based learning. *Brain*
751 *and Cognition*, *56*(2), 129–140.
752 Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal
753 and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397.
754 https://doi.org/10.1038/nn2066
755 Rushworth, M. F. S., Kolling, N., Sallet, J., & Mars, R. B. (2012). Valuation and decision-
756 making in frontal cortex: one or many serial or parallel systems? *Current Opinion in*
757 *Neurobiology*, *22*(6), 946–955.
758 Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action
759 sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–
760 417. https://doi.org/10.1016/j.tics.2004.07.009

761 Rypma, B., Berger, J. S., & D'Esposito, M. (2002). The Influence of Working-Memory
762 Demand and Subject Performance on Prefrontal Cortical Activity. *Journal of Cognitive*
763 *Neuroscience*, *14*(5), 721–731. https://doi.org/10.1162/08989290260138627
764 Salamone, J. D., Correa, M., Farrar, A., & Mingote, S. M. (2007). Effort-related functions of
765 nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology*,
766 *191*(3), 461–482. https://doi.org/10.1007/s00213-006-0668-9
767 Salamone, J. D., Correa, M., Yohn, S., Lopez Cruz, L., San Miguel, N., & Alatorre, L. (2016).
768 The pharmacology of effort-related choice behavior: Dopamine, depression, and
769 individual differences. *Behavioural Processes*, *127*, 3–17.
770 https://doi.org/10.1016/j.beproc.2016.02.008
771 Schmidt, L., Lebreton, M., Cléry-Melin, M.-L., Daunizeau, J., & Pessiglione, M. (2012).
772 Neural mechanisms underlying motivation of mental versus physical effort. *PLoS*
773 *Biology*, *10*(2), e1001266. https://doi.org/10.1371/journal.pbio.1001266
774 Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An
775 integrative theory of Anterior Cingulate Cortex function. *Neuron*, *79*(2), 217–240.
776 https://doi.org/10.1016/j.neuron.2013.07.007
777 Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the
778 value of control. *Nature Neuroscience*, *19*(10), 1286–1291.
779 https://doi.org/10.1038/nn.4384
780 Silvetti, M., Seurinck, R., & Verguts, T. (2011). Value and prediction error in medial frontal
781 cortex: integrating the single-unit and systems levels of analysis. *Frontiers in Human*
782 *Neuroscience*, *5*, 75. https://doi.org/10.3389/fnhum.2011.00075
783 Silvetti, M., Seurinck, R., & Verguts, T. (2013). Value and prediction error estimation account
784 for volatility effects in ACC: a model-based fMRI study. *Cortex*, *49*(6), 1627–1635.
785 https://doi.org/10.1016/j.cortex.2012.05.008

786 Silvia, P. J., Mironovová, Z., McHone, A. N., Sperry, S. H., Harper, K. L., Kwapil, T. R., &
787 Eddington, K. M. (2016). Do Depressive Symptoms "Blunt" Effort? An Analysis of
788 Cardiac Engagement and Withdrawal for an Increasingly Difficult Task. *Biological*
789 *Psychology*. https://doi.org/10.1016/j.biopsycho.2016.04.068
790 Silvia, P. J., Nusbaum, E. C., Eddington, K. M., Beaty, R. E., & Kwapil, T. R. (2014). Effort
791 Deficits and Depression: The Influence of Anhedonic Depressive Symptoms on Cardiac
792 Autonomic Activity During a Mental Challenge. *Motivation and Emotion*, *38*(6), 779–
793 789. https://doi.org/10.1007/s11031-014-9443-0
794 Treadway, M. T. (2016). The Neurobiology of Motivational Deficits in Depression--An Update
795 on Candidate Pathomechanisms. *Current Topics in Behavioral Neurosciences*, *27*, 337–
796 355. https://doi.org/10.1007/7854_2015_400
797 Treadway, M. T., Bossaller, N. A., Shelton, R. C., & Zald, D. H. (2012). Effort-based decision-
798 making in major depressive disorder: A translational model of motivational anhedonia.
799 *Journal of Abnormal Psychology*, *121*(3), 553–558. https://doi.org/10.1037/a0028813
800 Treadway, M. T., Buckholtz, J. W., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., …
801 Zald, D. H. (2012). Dopaminergic mechanisms of individual differences in human effort-
802 based decision-making. *Journal of Neuroscience*, *32*(18), 6170–6176.
803 https://doi.org/10.1523/JNEUROSCI.6459-11.2012
804 Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E., & Zald, D. H.
805 (2009). Worth the "EEfRT"? The effort expenditure for rewards task as an objective
806 measure of motivation and anhedonia. *PloS One*, *4*(8), e6598.
807 https://doi.org/10.1371/journal.pone.0006598
808 Treadway, M. T., Peterman, J. S., Zald, D. H., & Park, S. (2015). Impaired effort allocation in
809 patients with schizophrenia. *Schizophrenia Research*, *161*(2–3), 382–385.
810 https://doi.org/10.1016/j.schres.2014.11.024

van den Berg, B., Krebs, R. M., Lorist, M. M., & Woldorff, M. G. (2014). Utilization of reward-prospect enhances preparatory attention and reduces stimulus conflict. *Cognitive, Affective & Behavioral Neuroscience*, *14*(2), 561–577. https://doi.org/10.3758/s13415-014-0281-z

van Veen, V., Holroyd, C. B., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2004). Errors without conflict: implications for performance monitoring theories of anterior cingulate cortex. *Brain and Cognition*, *56*(2), 267–276. https://doi.org/10.1016/j.bandc.2004.06.007

Vassena, E., Cobbaert, S., Andres, M., Fias, W., & Verguts, T. (2015). Unsigned value prediction-error modulates the motor system in absence of choice. *NeuroImage*, *122*, 73–79. https://doi.org/10.1016/j.neuroimage.2015.07.081

Vassena, E., Holroyd, C. B., & Alexander, W. H. (2017). Computational models of anterior cingulate cortex: At the crossroads between prediction and effort. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00316

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

Vassena, E., Silvetti, M., Boehler, C. N., Achten, E., Fias, W., & Verguts, T. (2014). Overlapping neural systems represent cognitive effort and reward anticipation. *PLoS ONE*, *9*(3), e91008. https://doi.org/10.1371/journal.pone.0091008

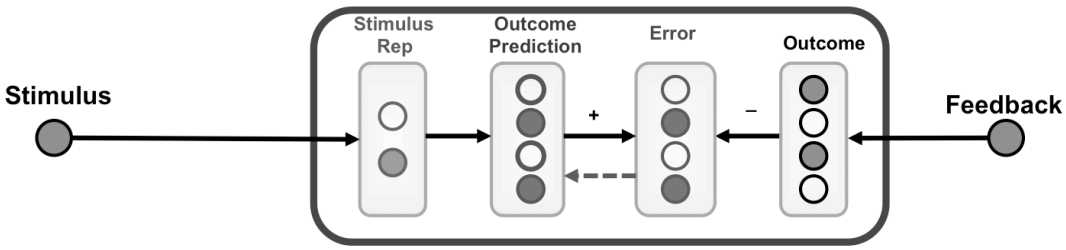
Verguts, T., Vassena, E., & Silvetti, M. (2015). Adaptive effort investment in cognitive and physical tasks: a neurocomputational model. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00057

834 Walton, M. E., Bannerman, D. M., Alterescu, K., & Rushworth, M. F. S. (2003). Functional
835 specialization within medial frontal cortex of the anterior cingulate for evaluating effort-
836 related decisions. *Journal of Neuroscience*, *23*(16), 6475–6479.
837 Walton, M. E., Bannerman, D. M., & Rushworth, M. F. S. (2002). The role of rat medial frontal
838 cortex in effort-based decision making. *Journal of Neuroscience*, *22*(24), 10996–11003.
839 Walton, M. E., Groves, J., Jennings, K. A., Croxson, P. L., Sharp, T., Rushworth, M. F. S., &
840 Bannerman, D. M. (2009). Comparing the role of the anterior cingulate cortex and 6-
841 hydroxydopamine nucleus accumbens lesions on operant effort-based decision making.
842 *European Journal of Neuroscience*, *29*(8), 1678–1691. https://doi.org/10.1111/j.1460-
843 9568.2009.06726.x
844 Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M., & Rushworth, M. F. S.
845 (2006). Weighing up the benefits of work: behavioral and neural analyses of effort-
846 related decision making. *Neural Networks*, *19*(8), 1302–1314.
847 https://doi.org/10.1016/j.neunet.2006.03.005
848 Walton, M. E., Rudebeck, P. H., Bannerman, D. M., & Rushworth, M. F. S. (2007). Calculating
849 the cost of acting in frontal cortex. *Annals of the New York Academy of Sciences*, *1104*,
850 340–356. https://doi.org/10.1196/annals.1390.009
851 Westbrook, A., & Braver, T. S. (2016). Dopamine Does Double Duty in Motivating Cognitive
852 Effort. *Neuron*, *89*(4), 695–710. https://doi.org/10.1016/j.neuron.2015.12.029
853 Westbrook, A., & Braver, T. S. (2013). The economics of cognitive effort. *Behavioral and*
854 *Brain Sciences*, *36*(6), 704–705. https://doi.org/10.1017/S0140525X13001179
855 Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach.
856 *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 395–415.
857 https://doi.org/10.3758/s13415-015-0334-y

Yang, X.-H., Huang, J., Zhu, C.-Y., Wang, Y.-F., Cheung, E. F. C., Chan, R. C. K., & Xie, G.-R. (2014). Motivational deficits in effort-based decision making in individuals with subsyndromal depression, first-episode and remitted depression patients. *Psychiatry Research*, *220*(3), 874–882. https://doi.org/10.1016/j.psychres.2014.08.056

883 **Figures and figures captions**

884



885

886 **Figure 1**. PRO model architecture (adapted from Alexander and Brown, 2011). The circles
887 outside the box represent environmental input (stimulus and feedback). The circles inside the
888 box represent units coding neural activity. Stimulus representations code environmental stimuli.
889 Depending on previous occurrence, certain stimuli predict certain outcomes, as coded in
890 outcome prediction units. Outcome units code real environmental outcome (feedback). A
891 comparison between outcome prediction units and outcome units results in an error signal
892 (discrepancy between predicted and actual outcome). This error signal feeds back into the
893 outcome prediction unit, to update such predictions.

894

895

896

897

898

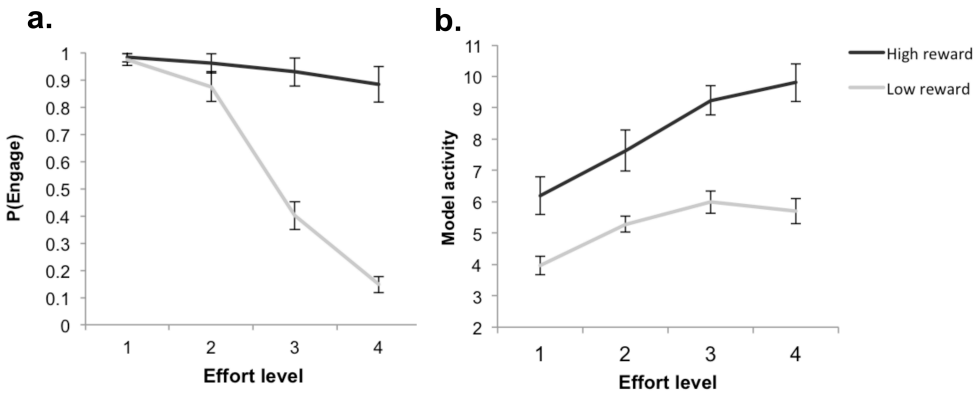


Figure 2. Model predictions. a. Behavioral predictions. The y-axis shows the probability of choosing to engage in a task. The x-axis shows four different effort levels (varying parametrically from easy (level 1) to hard (level 4). The grey line indicates a low reward upon successful completion. The black line indicates a high reward upon successful completion. The plot shows that with a low reward, increasing task difficulty reduces the probability of engaging in the task, while with a high reward the model engages in the increasingly effortful task anyway. b. Neural predictions. The y-axis shows MPFC activity at the time of cue. The x-axis shows the four effort levels. The grey line indicates low reward. The black line indicates high reward. The plot shows that model activity is overall higher when reward is high. Moreover, when reward is high activity linearly increases as a function of increasing effort. When reward is low, model activity only increases up to effort level 3.

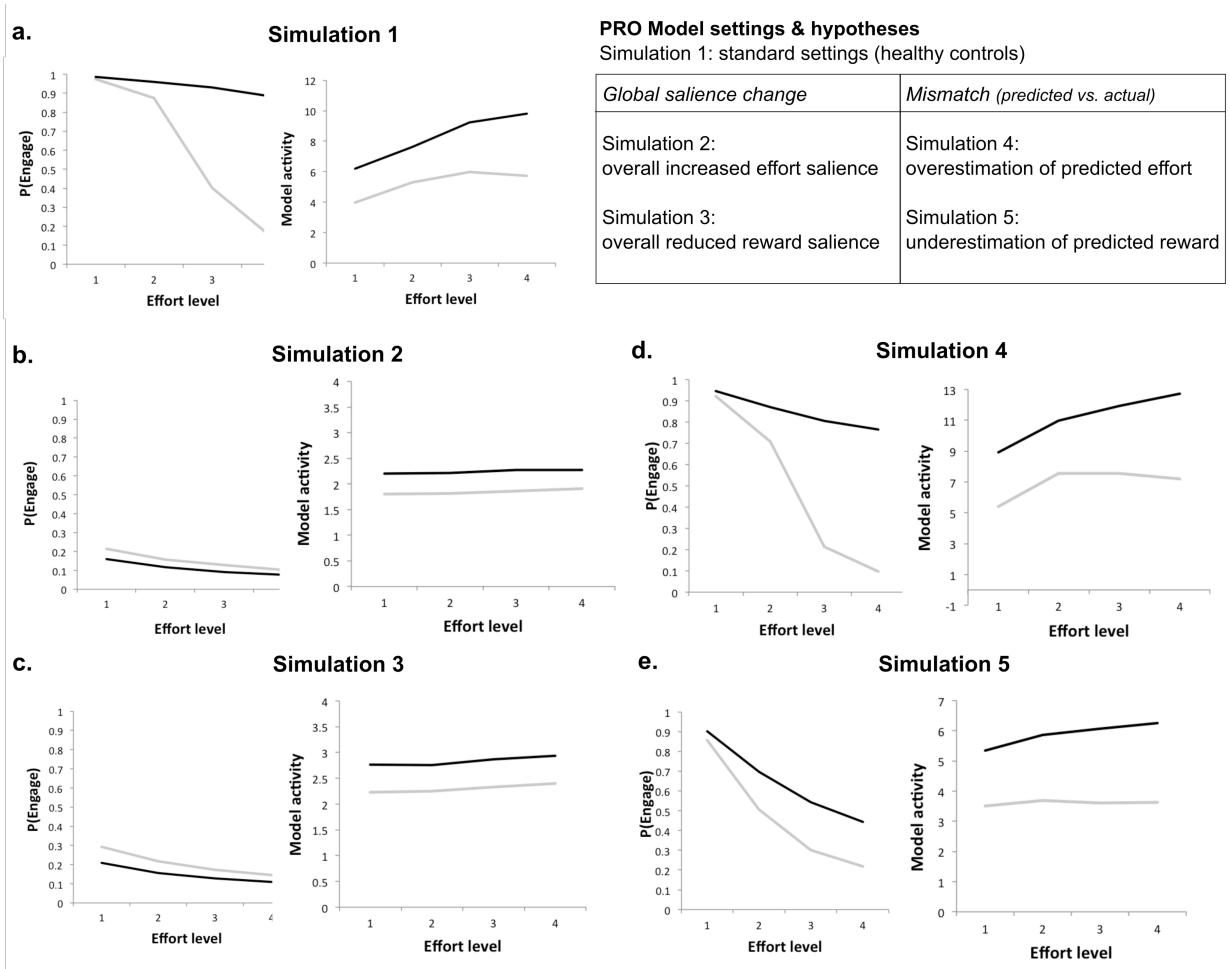
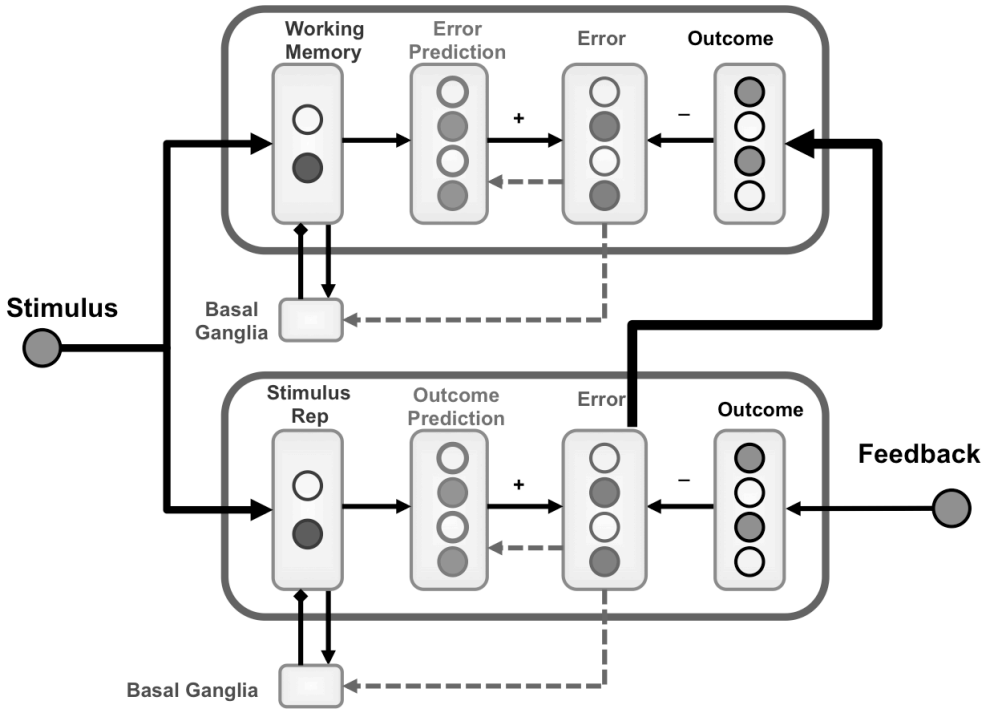


Figure 3. PRO model simulations of impaired motivation. In all plots, the y-axis shows the probability of engaging in the task (left panel) and the model activity (right panel). The x-axis shows four possible effort levels, parametrically increasing from easy (level 1) to hard (level 4). Grey lines indicate low reward upon task completion. Black lines indicate high reward upon task completion. a. Simulation 1. Behavioral and neural predictions for healthy controls. The table on the right illustrates the hypotheses of possible impairments as modeled with the PRO model, and relative explanation. We hypothesize two core possible mechanisms driving impairments in patients. The first is altered *global salience*, with either an overall increased effort salience (simulation 2), or an overall increased reward salience (simulation 3). The second is *mismatch* between predicted and actual outcome, with either a possible

924 overestimation of predicted effort (as compared to actual, simulation 4) or a possible
925 underestimation of the predicted reward (as opposed to actual, simulation 5). b-e. Simulations
926 2 through 5. Behavioral (choices) and neural (model activity) predictions under the different
927 hypotheses.

928



929

930 **Figure 4**. HER model architecture (adapted from Alexander and Brown, 2015). The circles
931 outside the box represent environmental input (stimulus and feedback). The circles inside the
932 box represent units coding neural activity. This figure shows a 2-layer version of the HER
933 model. Each layer replicates the architecture of the PRO model (cfr. Figure 1): stimulus
934 representation unit code for environmental stimuli, leading to prediction of a certain outcome
935 (outcome prediction units); outcome units code for actual outcome; comparison between
936 outcome prediction and actual outcome produces an error signal, coded in the error units. In the
937 HER model, the activity of error units at the lowest layer scales with the discrepancy between
938 predicted and actual outcome (as in the PRO model); error signals update predictions in this
939 layer, but are also fed to the upper layer, where actual error signal is compared with the

940 predicted error signal. The resulting 2nd level error signal is used to update predictions of the

941 future error signal.

942

943

944