Effects of depression on prefrontal–striatal goal-directed and habitual control

Suyeon Heo[1,2] and Sang Wan Lee*[1,2,3,4]

[1] Department of Bio and Brain Engineering,
[2] Brain and Cognitive Engineering Program,
[3] KAIST Institute for Health Science Technology,
[4] KAIST Institute for Artificial Intelligence,
Korea Advanced Institute of Science Technology (KAIST),
Daejeon 34141, Republic of Korea

*Corresponding author (sangwan@kaist.ac.kr)

ABSTRACT

Depression is characterized by deficits in the reinforcement learning (RL) process. Although many computational and neural studies have extended our knowledge of the impact of depression on RL, most focus on habitual control (model-free RL), yielding a relatively poor understanding of goal-directed control (model-based RL) and arbitration control to find a balance between the two. We investigate the effects of depression on goal-directed and habitual control in the prefrontal–striatal circuitry. We find that depression is associated with attenuated state and reward prediction error representation in the insula and caudate, a disruption of arbitration control in the predominantly inferior lateral prefrontal cortex and frontopolar cortex, and suboptimal value–action conversion. These findings fully characterize how depression influences different levels of RL, challenging previous conflicting views that depression simply influences either habitual or goal-directed control. Our study creates possibilities for various clinical applications, such as early diagnosis and behavioral therapy design.

INTRODUCTION

Major depressive disorder (MDD) has received considerable attention, as the lifetime prevalence of the disorder is higher than 10% worldwide[1]. MDD is characterized by deficits in

decision-making[2,3] and its underlying reward learning processes[4]. Recently, with the development of computational models, several studies have explored how depression influences the reward learning system (for a review, see (Chen et al., 2015)).

Reinforcement learning (RL), the process of learning to develop a behavioral policy to maximize reward[5], has been known to be guided by the two distinct RL strategies: model-based (MB) RL and model-free (MF) RL, each of which guides goal-directed and habitual RL, respectively[6]. Model-based RL guides context-sensitive and goal-directed behaviors through a sophisticated process in which the learning agent makes decisions by simulating an internal environmental model, whereas model-free RL is associated with habitual responses to reward-predicting stimuli based on learned associations between stimuli and rewards[6,7]. Mounting evidence suggests that depression is characterized by impairments in either model-based or model-free RL. For example, behaviors in depressive individuals can be accounted for by impaired model-based RL[8–10] or a transition from model-based to model-free RL[11]. However, most studies have only explored the effect of depression on model-free RL. For instance, depressive people exhibit an impaired ability to learn stimulus–reward associations accompanying inaccurate representations of reward prediction error[12–17] or abnormal learning rate control[11,18].

Impairment in RL is associated with not only the onset of depressive symptoms, but also the development of depression. For instance, stress, one of the major risk factors for depression[19,20], can induce deficits in RL. Previous findings have shown that people exhibit a reduced ability to engage in model-based RL under conditions of chronic[21,22] and acute[23,24] stress. These findings suggest a gradual impairment of RL from the very early stages of depression.

Although these studies have contributed to our understanding of depression in the context of RL, it is still unclear whether depression is best characterized by model-free RL, model-based RL, or an interaction between the two, or how depression influences the neural circuits guiding goal-directed and habitual behavioral control. Moreover, little is known about how these cases extend to early or mild depression.

Here, we aim to provide a computational and neural account of how depression affects goal-directed and habitual control in the prefrontal–striatal circuitry. First, to investigate the effects of depression on model-based and model-free RL, we ran a model comparison analysis to identify a version of arbitration control intended to account for various behavioral traits of depression. In particular, our computational models consider sub-optimality of RL, allowing us to explain choice behavior patterns across a wide spectrum of depression. We combine this

with model-based functional magnetic resonance imaging (fMRI) to identify the parametric effects of depression on neural systems associated with model-based and model-free RL. In the subsequent analysis, we attempt to fully characterize how depression disrupts the arbitration between model-based and model-free RL by combining the results from the computational modeling and model-based fMRI analyses and the multi-voxel pattern analysis (MVPA).

RESULTS

Effects of Depression on Behavior Performance of Goal-directed and Habitual Learning

[Figure 1] Markov decision task structure

Sixty-three participants conducted a sequential two-stage Markov decision task[25]. Of the subjects, 28 were scanned with an fMRI while performing the task. The task manipulated both prediction uncertainty and task goals to dissociate goal-directed and habitual behavior control (Figure 1; for more detail, see Methods). The task consisted of four types of blocks (high/low state–action–state transition uncertainty x specific/flexible goal condition). Before each experiment, participants completed the Center for Epidemiologic Studies Depression (CES-D) questionnaire[26] (For the distribution of participant's depression severity, see Supplementary Figure S2).

[Figure 2] Behavioral results

Overall task performance is negatively correlated with the self-reported depression score. The accumulated reward decreases significantly as individual depression score (CES-D) increases (correlation coefficient estimate=-0.584 [p=4.94e-07]; Figure 2a). The proportion of optimal choices, the measure that quantifies the extent to which a subject's choice reflects an optimal policy, is also inversely proportional to the CES-D score (correlation coefficient estimate=-0.567 [p=9.32e-05]; Figure 2b). Finally, choice consistency, the proportion of making the same choice as in previous trials, decreases as the CES-D score increases (correlation coefficient estimate=-0.472 [p=1.24e-06]; Figure 2c). These results demonstrate

that depression has a damaging effect on both goal-directed and habitual control performance, leading to suboptimal choices.

Computational Model of Arbitration Control Allowing for Suboptimal Decision-Making

[Figure 3] Computational model of dynamic arbitration control

We adopted the previous dynamic arbitration control hypothesis that respective prediction uncertainty—specifically, the amount of uncertainty in the state and reward prediction error of model-based and model-free RL—mediates the trial-by-trial value integration of the model-based and model-free systems[25]. To fully explore the effects of depression on arbitration control, however, a model should be flexible enough to account for any individual variability arising from suboptimal learning and decision-making.

To consider this, we redesigned the arbitration control scheme to allow for sub-optimality of RL in both learning values and converting learned values into choice behavior. We included the former by introducing separate learning rates for model-based and model-free RL and the latter by defining an exploitation sensitivity parameter as a function of model preference for either model-based or model-free RL (for more detail, see Methods). This model setting reflects the hypothesis that prediction uncertainty mediates not only value integration, but also value–action conversion (Figure 3a).

We compared prediction performance for the five different versions of arbitration control, including the original arbitration model and four other versions implementing our hypothesis in different ways. We used the Bayesian information criteria (BIC) as a performance measure to preclude overfitting. We found that the version implementing our hypothesis, in which the degree of exploitation is determined by the weighted sum of the model-based and model-free exploitation parameters with the model choice probability, explains the subjects' behavior ($t(62)=2.46$, [$p=0.017$]; paired t-test comparison with the second-best model, Arb_{α}), significantly better than the original model[25] (Figure 3b; for a model comparison, see Supplementary Table S1; for estimated model parameters, see Supplementary Table S2). Our model comparison result not only corroborates the previous finding that prediction uncertainty mediates the arbitration between model-based and model-free RL, but also demonstrates the effect of prediction uncertainty on both value integration and value–action conversion.

[Figure 4] Neural correlates of dynamic arbitration control

To further examine whether our model explains the neural activity patterns of brain areas previously implicated in model-based and model-free RL, we ran a model-based fMRI analysis in which each of the key signals of our computational model were regressed against the fMRI data.

First, we replicated previous findings concerning the neural correlates of prediction error for the model-based and model-free systems. The state prediction error (SPE) was found bilaterally in the insula and the dorsolateral prefrontal cortex (dlPFC) (all $p<0.05$ in cluster-level corrected). The reward prediction error (RPE) was correlated with neural activity in both sides of the ventral striatum ($p<0.05$ in family-wise error [FWE] corrected). These results are fully consistent with previous findings[25,27,28] (Figure 4, Supplementary Table S3).

We also successfully replicated previous findings supporting the neural hypothesis of arbitration control. We found the max reliability signal, the key signal used to mediate arbitration between model-based and model-free RL, in the bilateral inferior lateral PFC (ilPFC, left: $p<0.05$ in cluster-level corrected; right: $p<0.05$ in small-volume corrected [SVC] at [8,6,-2]) and the frontopolar cortex (FPC, $p<0.05$ in cluster-level corrected), fully consistent with previous results[25,29] (Figure 4, Supplementary Table S3).

Next, we tested the brain areas implicated in value computation. The chosen value of the model-based system (Q_{MB}) was found to be encoded in the precentral gyrus ($p<0.05$ in cluster-level corrected) and the orbital and medial PFC ($p<0.05$ in SVC at [-14,36,-8])[28,30]. The chosen value of the MF system (Q_{MF}) was found in the dorsal ACC, supplementary motor area, premotor cortex, dorsolateral PFC ($p<0.05$ in cluster-level corrected), and dorsomedial PFC ($p<0.05$ in SVC at [9,35,40])[31,32]. Notably, this model-free value signal was also found in the posterior putamen, the brain area known to be involved in valuation for habitual learning ($p<0.05$ in SVC at [-27,-4,1])[32–34]. We also tested for the integrated value signal, expressed as a sum of the value estimates of the model-based and model-free systems weighted by the arbitration control signal (P_{MB}). The ventromedial PFC was positively correlated with the difference between the integrated value signals for the chosen and unchosen actions ($p<0.05$ in cluster-level corrected), fully consistent with previous reports on choice values[35–37] (Figure

4, Supplementary Table S4).

Unlike the previous arbitration hypothesis[25], our computational model also predicted that arbitration control influences how integrated values are converted into actual choices. Finally, we attempted to identify the brain regions involved in value–action conversion. We found that the inferior parietal lobe, insula ($p<0.05$ in FWE corrected), middle frontal gyrus, globus pallidus, FPC, supplementary motor area, and thalamus ($p<0.05$ in cluster-level corrected) are positively correlated with the probability value of the chosen action, referred to as the output value of the softmax function[38]. This finding is consistent with previous findings indicating stochastic action selection[39]. Other brain areas, such as the orbitofrontal cortex, superior temporal gyrus, middle frontal gyrus, supramarginal gyrus ($p<0.05$ in FWE corrected), medial PFC, superior frontal gyrus ($p<0.05$ in cluster-level corrected), posterior medial cortex, and lateral PFC, are negatively correlated with the chosen action probability. The negative encoding of stochastic action selection in the posterior medial cortex and the lateral PFC also replicates previous findings, as these regions have been implicated in the valuation of counterfactual choices[35,39,40] (Figure 4, Supplementary Table S5).

Effects of Depression on Goal-directed and Habitual Learning

[Figure 5] Parametric effect of depression on goal-directed and habitual learning

To fully explore how depression affects the neural computations underlying model-based and model-free RL, we examined the relationship between the individual depression score and neural representations in each brain region implicated in model-based and model-free RL. We found evidence indicating the effect of depression on RL in multiple brain areas encoding prediction errors. The correlation coefficient between left insula activation and the SPE ([-36,20,-4], z=4.49), which represents the efficiency of neural encodings of the SPE in the left insula, was inversely proportional to the depression score (estimated correlation coefficient=-0.396, p=0.037; Figure 5a). We also found a significant negative correlation between the neural efficiency for encoding the RPE in the bilateral caudate (left: [-4,6,-4], z=4.50, right: [4,8,-4], z=5.30) and the individual depression score (estimated correlation coefficient=-0.412/-0.376, p=0.029/0.049 for the left and right, respectively; Figure 5b). These findings directly demonstrate how depression affects value updates for goal-directed and habitual learning.

Effects of Depression on Prefrontal Arbitration Control

[Figure 6] Parametric effect of depression on prefrontal arbitration control

We also tested for the neural effects of depression on arbitration control. First, we found that the individual depression score is significantly correlated with the learning rate for MF reliability estimation, the key variable required to quantify the reliability of predictions made by the model-free RL strategy based on the RPE (estimated correlation coefficient=0.335, p=0.007; Figure 6a, left; Supplementary Figure S1). This offers a theoretical prediction that depression entails over-sensitivity to the RPE, making arbitration control more sensitive to the prediction of the model-free system (Figure 6a, right). This implies that the brain regions implicated in mediating arbitration control focus on information about the reliability of the model-free system, rather than encoding the reliability information of the RL system controlling behavior at the moment (i.e. max reliability), leading to the disruption of neural computation underlying normal arbitration control.

To test the prediction that depression disrupts neural processing pertaining to arbitration control, we conducted a GLM analysis with the max reliability signal. We found that the effect size (parameter estimates from the GLM analysis) of the max reliability signal for FPC ([8,44,40], z=3.83) was negatively correlated with the depression score (estimated correlation coefficient=-0.441, p=0.019; Figure 6b). Moreover, the effect sizes of the max reliability signal for the bilateral ilPFC ([-52,26,16], z-score=4.48; [42,20,-8], z-score=3.61) and FPC ([8,44,40], z-score=3.83), the brain areas previously implicated in arbitration control[25,29,41], tended to be lower in the depressive group (CES-D score≥16) than in the control group (CES-D score<16) (Figure 6c, left; one-way ANOVA; $F_{1,26}$=2.76 [p=0.108], $F_{1,26}$=2.93 [p=0.099], $F_{1,26}$=5.18 [p=0.031] for the left and right ilPFC and FPC, respectively).

To further evaluate the prediction that depression makes neural processing for arbitration control more sensitive to the model-free system, we ran an MVPA for the bilateral ilPFC and FPC. This analysis quantifies the amount of information concerning the reliability of the model-free system embedded in these brain areas. We used a support vector machine, an optimal neural network for prediction and generalization, to conduct a binary classification of model-free reliability (high vs. low; upper/lower 33rd percentile threshold) and compared the prediction performance of the control and depression groups.

We found that the prediction performance of model-free reliability in the bilateral ilPFC and FPC was significantly higher in the depression group than in the control group (Figure 6c, right; one-way ANOVA; $F_{1,26}$=4.57 [p=0.042], $F_{1,26}$=4.33 [p=0.047], $F_{1,26}$=4.89 [p=0.036] for the left and right ilPFC and FPC, respectively). On the other hand, the same analysis found no significant inter-group differences in model-based reliability signal or max reliability signal (Supplementary Table S6). Taken together, these results strongly support our arbitration control hypothesis that depression is associated with increased sensitivity to RPE, leading to instable arbitration in which the reliability of predictions of the model-free system becomes predominant and the reliability of predictions of the model-based system becomes less influential.

Effects of Depression on Value–Action Conversion

[Figure 7] Parametric effect of depression on value-action conversion

Our computational model also explains how the degrees of control allocated to the model-based and model-free systems influence how value is converted into actual choice (exploitation sensitivity). For example, exploitative and explorative choices are associated with high and low exploitation sensitivity, respectively. We found that the exploitation parameter for model-based RL is negatively correlated with the individual depression score (correlation coefficient=-0.412, p=0.001; Figure 7a, left), indicating that subjects with higher depression scores exhibit more exploratory choices when their choices are guided by model-based RL (Figure 7a, right).

In the subsequent neural analysis, we explored the relationships between the depression score and the neural representations of value–action conversion. The parameter estimates of the probability value of taking the chosen action are significantly correlated with the depression score in two brain regions. The parameter estimates in the two seed regions—the right superior temporal gyrus (STG; [-46,-20,-8], z=4.21) and left middle temporal gyrus (MTG; [56,-32,-8], z=3.55)—are positively and negatively correlated with the CES-D score, respectively (correlation coefficient=0.430, p=0.025 for STG; correlation coefficient=-0.430, p=0.022 for MTG; Figure 7b).

DISCUSSION

By combining a computational model allowing for sub-optimality in learning and decision-making, a model-based fMRI analysis, and an MVPA, the present study fully characterizes how depression influences the different levels and stages of RL: value computation, prefrontal arbitration control for value integration, and value–action conversion. We found that depression has a parametric effect on neural representations of prediction error for model-based and model-free systems, respectively, explaining how depression hampers value computation ability. Another intriguing finding is that the brain areas implicated in arbitration control, bilateral ilPFC and FPC, become more sensitive to the predictions of the model-free system in people with depression, indicating that depression disrupts the balance between goal-directed and habitual control. We also found that depression increases the tendency to make exploratory choices during model-based control, but not during model-free control.

Computational Theory of Prefrontal Goal-directed and Habitual Control

The present study's computational model of dynamic arbitration control of model-based and model-free RL allows us to explore the full parametric effects of depression on prefrontal goal-directed and habitual control. Although mounting evidence suggests that prediction uncertainty might be a key variable for prefrontal arbitration control[6,25,27], little is known about the computational reasons people with depression tend to exhibit behavioral biases towards either goal-directed or habitual behavior. Addressing this issue involves a few challenges. First, simply evaluating the two separate hypotheses contradicts the prevailing view that the brain circuitries guiding goal-directed and habitual behavior interact with each other. Second, there is no guarantee that a rational arbitration control model is flexible enough to explain the individual variability associated with depression. Third, exploring a depression-specific model based on the assumption that depression follows a computational regime that substantially deviates from rational decision-making may enable us to explain severe depression, but cannot explain a continuum extending from a normal to a severely depressed state.

To fully address these issues, we considered a computational model of dynamic arbitration control allowing for individual variability in suboptimal learning and decision making. Intriguingly, we found that the influence of prediction uncertainty is not confined to value integration[25], but extends as far as value–action conversion. This also allowed us to test the full effect of depression on decision-making at different computation levels: model-based and model-free reinforcement learning, arbitration control for value integration and value–action conversion.

Effects of Depression on Neural Representations of Prediction Error

Our study found that depression has a parametric effect on the neural representations of the two distinct types of prediction errors associated with model-based and model-free RL: SPE and RPE. The neural analysis revealed that depression scores were correlated with an attenuation of the SPE signal in the left insula and the RPE in the bilateral caudate.

Dopamine is crucial for both model-free and model-based RL. Numerous previous studies have reported dopamine's role in guiding the RPE[42,43], and a recent finding discussed the essential role of dopamine in stimulus-stimulus associative learning[44], implicating the involvement of dopamine in SPE representation. Depression is characterized by decreased dopamine levels[45,46], which may impair learning in both model-free and model-based systems. In fact, RPE signals in depression have reportedly been reduced in various experimental conditions (both Pavlovian learning[12] and instrumental learning[13–17]). Our study not only corroborates previous findings concerning RPE deficits in depression, but also further suggests that depression may impact neural representations of SPE.

Effects of Depression on Prefrontal Arbitration Control

One interesting prediction of the model is that CES-D is positively correlated with the parameter value for controlling the learning rate for updating model-free reliability based on the RPE, indicating that reliability estimation for the model-free strategy is very sensitive to RPE changes. This suggests that, in people with high CES-D scores, arbitration control might be predominantly driven by the model-free reliability signal, rather than by a fair comparison of the model-free and model-based reliability signals. We explored this possibility through a combination of a general linear model (GLM) and MVPA.

Our GLM analysis showed the negative effect of the depression on the neural representations of arbitration control in the prefrontal cortex. In bilateral inferior lateral PFC and frontopolar cortex, the brain areas reportedly encoding the key variable for arbitration control[25], neural representations tend to be weaker in the high CES-D score group. Critically, the subsequent MVPA shows that the amount of reliability information of the model-free system is significantly higher in the high CES-D score group. Taken together, these findings theoretically implicate that depression may hinder the PFC's ability to estimate the reliability of each learning strategy from the corresponding prediction error.

Effects of Depression on Valuation–Action Conversion

The present study also provides a computational and neural account of how depression causes sub-optimal action selection. Our computational model predicts that depression increases the tendency to make exploratory choices during model-based control, rather than model-free control.

This finding could also clarify the two conflicting views of choice consistency behaviors in depression[47,48]. Beever et al. (2013) found no significant difference in exploration pattern in a reward-maximizing task between a normal and a depressed group. Blanco et al. (2013), on the other hand, found that a depression group tended to explore more. This conflict might be attributable to differences in task structure. Beever et al.'s study used a task with a relatively stable environmental structure, such that people performed tasks relying on the model-free system. This is consistent with our view that depression has a relatively weak influence on exploration during model-free control. However, the reward structure used in Blanco's study encouraged more frequent policy changes, accommodating the need for model-based control. This is also consistent with our view that exploratory choice behavior becomes more pronounced during model-based control.

The neural results of the present study, which show that the STG response is higher in people with depression, are fully consistent with previous finding that STG response increases when people switch to other options rather staying[49]. Our results address not only the implication for the role of STG in exploration, but also how depression influences exploration at the neural level.

Our finding that the degree of exploitation decreases as CES-D score increases (shown in Figure 7) explains why reward sensitivity is reduced in people with depression. A decreasing degree of exploitation decreases the tendency to convert a learned policy into an actual choice, reducing the efficiency of translating changes in the reward structure into changes in actual choices. This is also consistent with the view that our model's degree of exploitation parameter can be interpreted as reward sensitivity[16,50]. In addition to supporting existing evidence of declined reward sensitivity in depression[51–53], the present study advances the view by proposing that this tendency becomes stronger during model-based control.

Potential clinical applications

The present findings suggest how depression influences goal-directed and habitual control in the prefrontal–striatal circuitry. The full characterization of the effects of depression on

different stages of learning and decision-making creates possibilities for various clinical applications. First, our neural results explain why such brain stimulus techniques as repetitive transcranial magnetic stimulation (rTMS) and deep brain stimulus (DBS) to the frontopolor cortex[54] or ventral striatum[55] are effective in alleviating depressive symptoms. Second, our theoretical idea suggests that behavioral therapy to reduce sensitivity to reward prediction errors might help people with depression regain a balance between goal-directed and habitual control. Intriguingly, our findings also indicate a parametric effect of depression on learning and decision-making in a relatively young age group (average=22.8 yrs). Considering the onset of MDD is approximately 25 to 45 yrs[56], our study offers possibilities for not only investigating how mild depression transitions to MDD, but also developing clinical applications for the early diagnosis of MDD.

METHODS

Participants

Sixty-five right-handed Koreans (28 females; mean age of 22.8±3.8) participated in the study. Participants were recruited from the local society through the online announcement. Only 28 subjects were scanned with fMRI during the task. Two subjects whose total accumulated reward are below the chance-level (mean amount of rewards with 10,000 random simulations) were excluded from the analysis. Thus, a total of sixty-three behavioral data and twenty-eight fMRI data were left for the analysis. To acquire the depressive level of individuals, people were instructed to complete the Center for Epidemiologic Studies Depression (CES-D) questionnaires[26] before the experiment (For the distribution of participant's depression severity, see Supplementary Figure S2).

No subjects had the history of neurological of psychiatric disease. Every subject provided written consent to the experimental protocols which were approved by the Institutional Review Board (IRB) of the Korea Advanced Institute of Science and Technology (KAIST).

Task

We used the sequential two-stage Markov decision task proposed to dissociate model-based and model-free learning strategy[25]. In this task, subjects make a binary choice (either left or right) and proceed to the next state with a certain probability. When the next state is appeared in the screen, participants make another choice. The two consecutive choices is followed by a transition to an outcome state. Subjects perform 100 trials in the pre-learning session to

learn the structure of the task. Four main sessions with 80 trials on average follow the pre-learning session. Participants are instructed to collect as many coins as possible in the main sessions.

The task consists of two conditions: a specific-goal condition and flexible-goal condition. The goal condition is indicated by a color of a box at the beginning of each trial. In the specific-goal condition, participants are presented with a box with a specific color (red, blue, yellow). A monetary reward is given only when the coin color matches with the color of the given box. In the flexible-goal condition, on the other hand, participants are given a white coin box with which all types of coins become redeemable. Two types of state-transition probability are used to control the uncertainty of the environment. The state-transition probability (0.5, 0.5) and (0.9, 0.1) is intended to implement the highly-uncertain and relatively less uncertain environment, respectively. The four types of block (2 goal-conditions x 2 uncertainty conditions) are presented in pseudo-random order. Each block consists of 4-6 trials.

Computational Models

The computational model of this study is motivated by the previous arbitration control hypothesis that prediction uncertainty of the model-based and model-free RL is a key variable to guide value integration of the two corresponding systems[25]. The model consists of the three processes: value learning, arbitration, and action selection (Figure 3a).

In the value learning stage, both a model-based and model-free system learn action values for each state. A model-based system uses state prediction error (SPE = 1-expected transition probability) to update the state-action-state transition probability, by using a FORWARD learning[27] and learns action values by combining the learned state-action-state transition probability and reward in the outcome state. For a model-free system, on the other hand, the state-action value learning is based on RPE (RPE = actual value-expected value). It is implemented with a SARSA algorithm[5].

In the arbitration process, the reliability estimation of the model-based system was implemented with a hierarchical empirical Bayes method using the history of the SPE, the reliability of the model-free system was implemented with the Pearce-hall associability rule using an unsigned RPE. These estimated reliability signal were then used to guide the competition between the two systems, which is implemented with a dynamic two-state transition model. The output of this model is a model choice probability (P_{MB}), used as the control weight for value integration of the two systems. Finally, in the action selection stage,

the model selects the action stochastically using softmax rule[38]. For more details, refer Lee et al (2014).

In this study, we suggested two variants of arbitration control, allowing for sub-optimality in value learning, arbitration, and action selection: one version with separate model-based and model-free learning and another version with a dynamic exploitation. The former type of the model assumes the different learning rates of a model-based and a model-free system. The latter class of models is based on the former model, with the further assumption that the degree of exploitation, an indicator of optimality of the RL agent's policy, is a function of the model choice probability, P_{MB}. We tested three different types of exploitation as follows: logistics, linear, weighted linear. Note that in all cases, we set the parameters of the model in a way that is reduced to the original RL with a single exploitation parameter.

fMRI Data Acquisition

Functional imaging was conducted on a 3T Siemens (Magnetom) Verio scanner located in the KAIST brain imaging center (Daejeon). Forty-two axial slices were acquired with interleaved-ascending order at the resolution of 3 mm x 3 mm x 3 mm, covering the whole brain. A one-shot echo-planner imaging pulse sequence was utilized (TR = 2800 ms; TE = 30 ms; FOV = 192 mm; flip angle = 90º). The high resolution structural image was also acquired for each subject to the resolution of 0.7 mm X 0.7 mm x 0.7 mm.

fMRI Data Pre-processing

Images were processed and analyzed using the SPM12 software (Wellcome Department of Imaging Neuroscience, London, UK). The first two volumes were removed to reduce T1 equilibrium effects. The EPI images were corrected for slice timing, motion movement and spatially normalized to the standard template imaging provided by SPM software.

For the general linear model analysis (GLM), normalized images were smoothed with 6mm FWHM Gaussian Kernel and a high-pass filter (128s cut-off) was applied to remove the noise.

For the multivoxel pattern analysis (MVPA), unsmoothed EPI image data was used. De-trending and z-scoring were processed to reduce the linear trends and to match the range of the signal.

General Linear Model Analysis (GLM)

Subject-specific value-related signals and arbitration control signals were computed from the arbitration model, and the signals were regressed against voxel-wise signals from the EPI image set. The order of the regressors is as follows: prediction error from the model-based system (SPE) and the model-free system (RPE), reliability comparison signal which is a key variable for arbitration control (=max (Rel_{MB}, Rel_{MF}); max reliability), the chosen value of model-based system (Q_{fwd}), the chosen value of model-free system (Q_{sarsa}), the difference between chosen and unchosen integrated values (Q_{arb}) and the probability of selecting chosen action ($P_{chosen\ action}$). The regressors were serially non-orthogonalized in the GLM analysis to prevent the effect of regressor orders in the interpretation of the results. MARSBAR software (http://marsbar.sourceforge.net) was used to extract parameter estimates from the region of interest[57].

Multivoxel Pattern Analysis (MVPA)

The MVPA analysis was conducted to quantify types and amounts of information encoded in specific the region of interest (ROI). The classification performance is regarded as the amount of information pertaining to the variable of interest. Three ROIs, left/right inferior lateral prefrontal cortex (ilPFC) and Frontopolar prefrontal cortex (FPC), were selected, which were known to engage in the arbitration control process. Masks of each brain region were functionally defined from the GLM analysis. We used the clusters whose response to the Max reliability signal survived after the whole-brain correction (p<0.001, uncorrelated) as a mask for each ROI. We set the BOLD response time 4-6s.

A binary Support Vector Machine (SVM) classifier was applied to learn voxels patterns with each ROI. For each subjects data, the SVM was trained to best match its output to a binarized reliability-related signal (MB reliability, MF reliability, or Max reliability); the 33th and 67th percentile threshold were used to define the two classes, 'high value group' and 'low value group', respectively. All voxels in the mask were used for learning. The input dimension is 196, 79, 294 for left ilPFC, right ilPFC and FPC, respectively. The average number of data from each subject are 350 for MB reliability, 549 for MF reliability and 543 for Max Reliability. Thirty-fold cross validation was conducted for evaluation. All processes were implemented based on the Princeton Multi-Voxel Pattern Analysis toolbox[58]. Finally, an ANOVA analysis was conducted to compare signal prediction accuracy between the normal and depression group; the two subject groups were defined by using the standard cutoff criteria of CES-D score, 16[59,60].

Data Availability

All data analyzed in this study will be available upon request.

ACKNOWLEDGEMENTS

This research was supported by the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2016M3C7A1914448), NRF funded by the Korea government (MSIT) (No. NRF-2017R1C 1B 2008972), the research fund of the KAIST (Korea Advanced Institute of Science and Technology) (Grant code: G04150045), and Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451).

Author Contributions

S.H. and S.W.L. designed the study, analyzed the data and wrote the manuscript. S.H. conducted the experiments.

Competing interests

No conflict of interests.

References

1. Lim, G. Y. *et al.* Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. *Sci. Rep.* **8,** 2861 (2018).

2. Leykin, Y., Roberts, C. S. & DeRubeis, R. J. Decision-Making and Depressive Symptomatology. *Cognit. Ther. Res.* **35,** 333–341 (2011).

3. Must, A., Horvath, S., Nemeth, V. L. & Janka, Z. The Iowa Gambling Task in depression - what have we learned about sub-optimal decision-making strategies? *Front. Psychol.* **4,** 732 (2013).

4. Admon, R. & Pizzagalli, D. A. Dysfunctional Reward Processing in Depression. *Curr. Opin. Psychol.* **4,** 114–118 (2015).

5. Sutton, R. S. & Barto, A. G. *Introduction to Reinforcement Learning*. (MIT Press, 1998).

6. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8,** 1704–1711 (2005).

7. Doya, K., Samejima, K., Katagiri, K. & Kawato, M. Multiple Model-Based Reinforcement Learning. *Neural Comput.* **14,** 1347–1369 (2002).

8. Markman, K. & Miller, A. *Depression, Control, and Counterfactual Thinking: Functional for Whom? Journal of Social and Clinical Psychology* **25,** (2006).

9. Quelhas, A. C., Power, M. J., Juhos, C. & Senos, J. Counterfactual thinking and functional differences in depression. *Clin. Psychol. Psychother.* **15,** 352–365 (2008).

10. Huys, Q. J. M., Daw, N. D. & Dayan, P. Depression: a decision-theoretic analysis. *Annu. Rev. Neurosci.* **38,** 1–23 (2015).

11. Maddox, W. T. *et al.* Elevated Depressive Symptoms Enhance Reflexive but not Reflective Auditory Category Learning. *Cortex.* **58,** 186–198 (2014).

12. Kumar, P. *et al.* Abnormal temporal difference reward-learning signals in major depression. *Brain* **131,** 2084–2093 (2008).

13. Gradin, V. B. *et al.* Expected value and prediction error abnormalities in depression and schizophrenia. *Brain* **134,** 1751–1764 (2011).

14. Dombrovski, A. Y., Szanto, K., Clark, L., Reynolds, C. F. & Siegle, G. J. Reward signals, attempted suicide, and impulsivity in late-life depression. *JAMA psychiatry* **70,** 1 (2013).

15. Ubl, B. *et al.* Altered neural reward and loss processing and prediction error signalling in depression. *Soc. Cogn. Affect. Neurosci.* **10,** 1102–1112 (2015).

16. Rothkirch, M., Tonn, J., Kohler, S. & Sterzer, P. Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain* **140,** 1147–1157 (2017).

17. Kumar, P. *et al.* Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. *Neuropsychopharmacology* **43,** 1581–1588 (2018).

18. Chase, H. W. *et al.* Approach and avoidance learning in patients with major

depression and healthy controls: relation to anhedonia. *Psychol. Med.* **40,** 433–440 (2010).

19. Plieger, T., Melchers, M., Montag, C., Meermann, R. & Reuter, M. Life stress as potential risk factor for depression and burnout. *Burn. Res.* **2,** 19–24 (2015).

20. Radley, J. J. *et al.* STRESS RISK FACTORS AND STRESS-RELATED PATHOLOGY: NEUROPLASTICITY, EPIGENETICS AND ENDOPHENOTYPES. *Stress* **14,** 481–497 (2011).

21. Dias-Ferreira, E. *et al.* Chronic stress causes frontostriatal reorganization and affects decision-making. *Science* **325,** 621–625 (2009).

22. Radenbach, C. *et al.* The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology* **53,** 268–280 (2015).

23. Schwabe, L. & Wolf, O. T. Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. *Behav. Brain Res.* **219,** 321–328 (2011).

24. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl. Acad. Sci.* **110,** 20941 LP-20946 (2013).

25. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81,** 687–699 (2014).

26. Radloff, L. S. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Appl. Psychol. Meas.* **1,** 385–401 (1977).

27. Glascher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66,** 585–595 (2010).

28. O'Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, Reward, and Decision Making. *Annu. Rev. Psychol.* **68,** 73–100 (2017).

29. Bogdanov, M., Timmermann, J. E., Gläscher, J., Hummel, F. C. & Schwabe, L. Causal role of the inferolateral prefrontal cortex in balancing goal-directed and habitual control of behavior. *Sci. Rep.* **8,** 9382 (2018).

30. Wunderlich, K., Dayan, P. & Dolan, R. J. Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci.* **15,** 786–791 (2012).

31. Rowe, J. B., Hughes, L. & Nimmo-Smith, I. Action selection: A race model for selected and non-selected actions distinguishes the contribution of premotor and prefrontal areas. *Neuroimage* **51,** 888–896 (2010).

32. Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P. & Rangel, A. Transformation of stimulus value signals into motor commands during simple choice. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 18120–18125 (2011).

33. Tricomi, E., Balleine, B. W. & O'Doherty, J. P. A specific role for posterior dorsolateral striatum in human habit learning. *Eur. J. Neurosci.* **29,** 2225–2232 (2009).

34. Horga, G. *et al.* Changes in corticostriatal connectivity during reinforcement learning in humans. *Hum. Brain Mapp.* **36,** 793–803 (2015).

35. Boorman, E. D., Behrens, T. E. J., Woolrich, M. W. & Rushworth, M. F. S. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62,** 733–743 (2009).

36. Jocham, G., Hunt, L. T., Near, J. & Behrens, T. E. J. A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. *Nat. Neurosci.* **15,** 960 (2012).

37. Jocham, G. *et al.* Dissociable contributions of ventromedial prefrontal and posterior parietal cortex to value-guided choice. *Neuroimage* **100,** 498–506 (2014).

38. R Luce, D. *Individual Choice Behavior: A Theoretical Analysis*. *New York* **115,** (2005).

39. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441,** 876–879 (2006).

40. Boorman, E. D., Behrens, T. E. & Rushworth, M. F. Counterfactual choice and learning in a Neural Network centered on human lateral frontopolar cortex. *PLoS Biol.* **9,** (2011).

41. Cole, M. W., Repovs, G. & Anticevic, A. The frontoparietal control system: a central role in mental health. *Neuroscientist* **20,** 652–664 (2014).

42. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275,** 1593–1599 (1997).

43. Hollerman, J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1,** 304–309 (1998).

44. Sharpe, M. J. *et al.* Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20,** 735–742 (2017).

45. Dunlop, B. W. & Nemeroff, C. B. The role of dopamine in the pathophysiology of depression. *Arch. Gen. Psychiatry* **64,** 327–337 (2007).

46. Malhi, G. S. & Berk, M. Does dopamine dysfunction drive depression? *Acta Psychiatr. Scand.* **115,** 116–124 (2007).

47. Blanco, N. J., Otto, A. R., Maddox, W. T., Beevers, C. G. & Love, B. C. The influence of depression symptoms on exploratory decision-making. *Cognition* **129,** 563–568 (2013).

48. Beevers, C. G. *et al.* Influence of depression symptoms on history-independent reward and punishment processing. *Psychiatry Res.* **207,** 53–60 (2013).

49. Paulus, M. P., Feinstein, J. S., Leland, D. & Simmons, A. N. Superior temporal gyrus and insula provide response and outcome-dependent information during assessment and action selection in a decision-making situation. *Neuroimage* **25,** 607–615 (2005).

50. Huys, Q. J. M., Pizzagalli, D. A., Bogdan, R. & Dayan, P. Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biol. Mood Anxiety Disord.* **3,** 12 (2013).

51. Steele, J. D., Kumar, P. & Ebmeier, K. P. Blunted response to feedback information in depressive illness. *Brain* **130,** 2367–2374 (2007).

52. Chen, C., Takahashi, T., Nakagawa, S., Inoue, T. & Kusumi, I. Reinforcement learning in depression: A review of computational research. *Neurosci. Biobehav. Rev.* **55,** 247–267 (2015).

53. Alloy, L. B., Olino, T., Freed, R. D. & Nusslock, R. Role of Reward Sensitivity and Processing in Major Depressive and Bipolar Spectrum Disorders. *Behav. Ther.* **47,** 600–621 (2016).

54. Downar, J. & Daskalakis, Z. J. New Targets for rTMS in Depression: A Review of Convergent Evidence. *Brain Stimul. Basic, Transl. Clin. Res. Neuromodulation* **6,** 231–240 (2013).

55. Delaloye, S. & Holtzheimer, P. E. Deep brain stimulation in the treatment of depression. *Dialogues Clin. Neurosci.* **16,** 83–91 (2014).

56. Kessler, R. C. *et al.* Age of onset of mental disorders: a review of recent literature.

Curr. Opin. Psychiatry **20,** 359–364 (2007).

57. Brett, M., Anton, J.-L. L., Valabregue, R. & Poline, J.-B. Region of interest analysis using an SPM toolbox - Abstract Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. *Neuroimage* **16,** Abstract 497 (2002).

58. Detre, G. J. *et al.* The Multi-Voxel Pattern Analysis (MVPA) toolbox. *Ohbm* (2006).

59. Comstock, G. W. & Helsing, K. J. Symptoms of depression in two communities. *Psychol. Med.* **6,** 551–563 (1976).

60. Weissman, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A. & Locke, B. Z. Assessing depressive symptoms in five psychiatric populations: a validation study. *Am. J. Epidemiol.* **106,** 203–214 (1977).

Figures

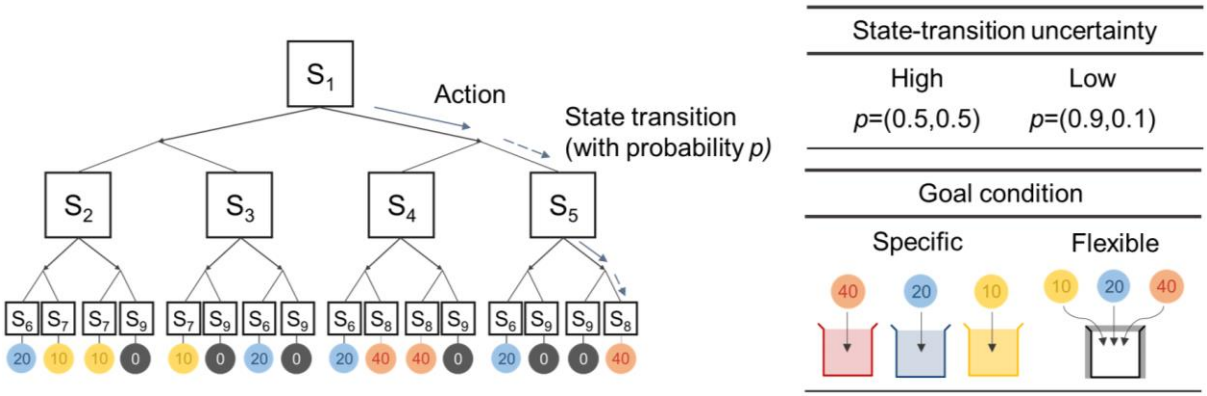

Figure 1. Markov decision task structure

We use the two-stage Markov decision task proposed by Lee et al. (2014). In each stage, participants make a binary choice (left or right). After the first choice in the initial state (S_1), they were moved forward to one of four states in the second stage (S_2 - S_5) with certain state-action-state transition probability *p*. The transition probability (0.5, 0.5) and (0.9, 0.1) corresponds to a high-uncertainty and a low-uncertainty environment, respectively. The task consists of the two goal conditions: a specific-goal and a flexible-goal condition. In the specific goal condition, subjects can collect coins (redeemable for monetary reward) only if the coin color matches with the color of the token box (red, blue, yellow). In the flexible goal condition indicated by the white token box, all types of coins are redeemable.

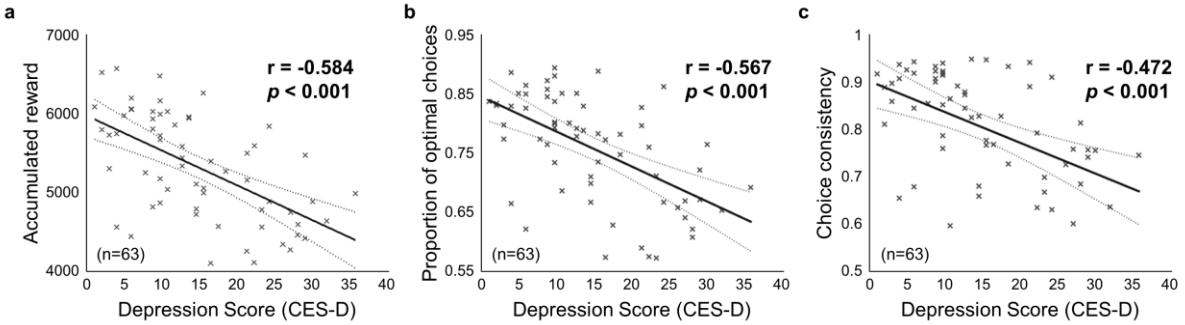

Figure 2. Behavioral results

(a) Relationship between the individual depression score (CES-D) and accumulated reward (n=63). The task performance decreases as the depression score increases. (b) Relationship between depression score and the proportion of optimal choices (n=63). The proportion of optimal choices is inversely proportional to the depression score. (c) Relationship between the depression score and choice consistency in the first state (S_1) (n=63). The choice consistency index is negatively correlated with the depression score.

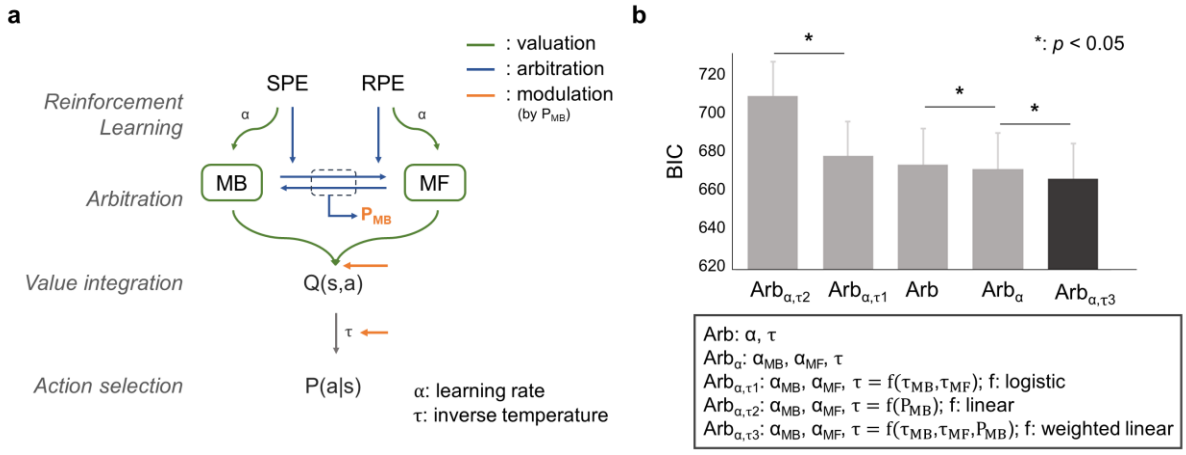

Figure 3. Computational model of dynamic arbitration control

(a) Computational model to investigate dynamic control mechanisms to arbitrate between model-based (MB) and model-free (MF) RL. Two separate RL systems update an action value and reliability of its prediction by using prediction errors (SPE and RPE, respectively; green). The reliability values of the two systems were then used to compute the model choice probability (P_{MB}) (blue). The model choice probability guides both the value integration and value-action conversion process; both effects are indicated by the orange arrow. (b) Model comparison analysis. We used Bayesian Information Criteria (BIC) for comparing the goodness of fit while penalizing for the model complexity. Arb refers to the computational model proposed in (Lee et al., 2014). Additional versions of arbitration control consider separate learning rates and dynamic exploitation. $Arb_{α}$ assigns separate learning rates to two different systems (i.e. use $α_{MB}$, $α_{MF}$ instead of α). $Arb_{α,τ}$ is the same as $Arb_{α}$, except for the assumption that the degree of exploitation is a function of the model choice probability. The best version of model uses the degree of exploitation (τ) as a weighted sum of the model-based and the model-free exploitation parameter ($τ_{MB}$ and $τ_{MF}$, respectively) with the model choice probability P_{MB} (i.e. $τ = P_{MB} * τ_{MB} + (1-P_{MB}) * τ_{MF}$). The error bar stands for the standard error mean.

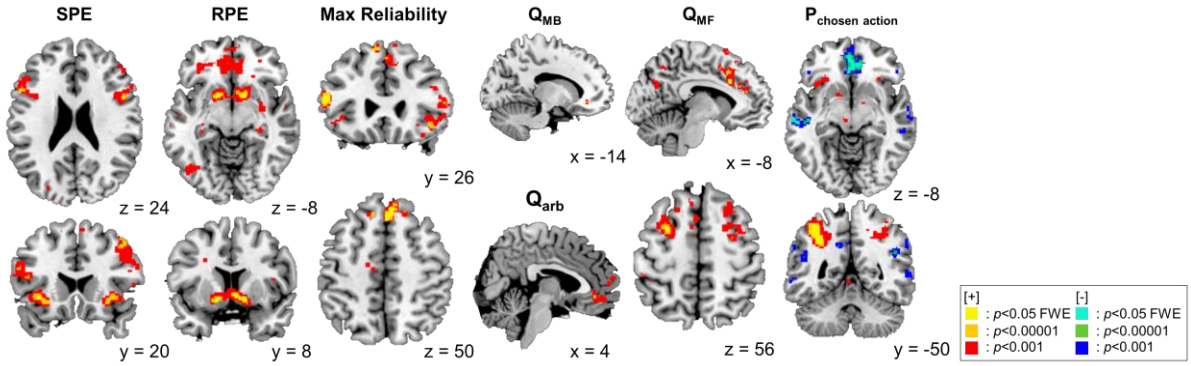

Figure 4. Neural correlates of dynamic arbitration control

Prediction error, value, reliability, action choice probability signals from the proposed model are shown as colored blobs. SPE and RPE refers to state prediction error and reward prediction error, respectively. Max reliability refers to the reliability of whichever system had the highest reliability index on each trial (=max(Rel_{MB}, Rel_{MF})). Q_{MB} and Q_{MF} indicate the chosen value from the model-based and model-free system. Q_{arb} refers to the difference between integrated chosen action value and unchosen action value. $P_{chosen\ action}$ refers to the probability assigned to the chosen action. See more detailed information in Supplementary Table S3-S5.

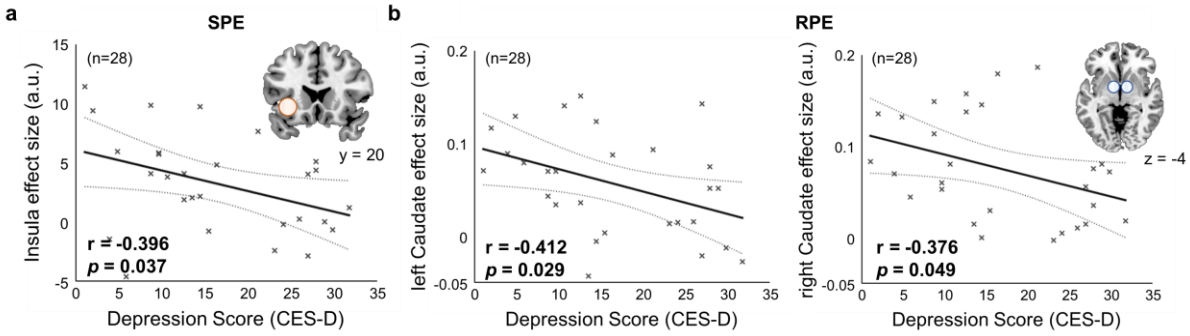

Figure 5. Parametric effect of depression on goal-directed and habitual learning

(a) Depression impacts on neural encoding of SPE information (n=28). The shaded circles represent seed regions for which parameter estimates of the GLM analysis were extracted. The seed region is the left insula, and the parameter estimates were extracted from our GLM analysis which regressed the SPE signal against the BOLD response ([-36,20,-4], z-score=4.49; Figure 4). The estimated effect size for the left insula is negatively correlated with the depression score. (b) Depression impacts on RPE response (n=28). The estimated effect size of RPE for bilateral caudate (left: [-4,6,-4], z-score=4.50, right: [4,8,-4], z-score=5.30) is negatively correlated with the depression score. a.u. stands for arbitrary units.

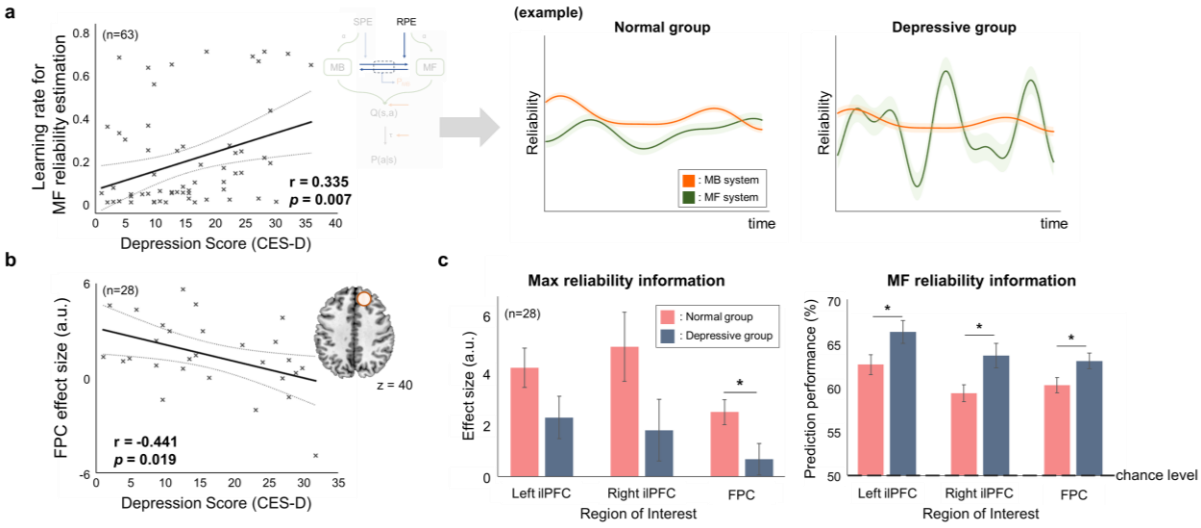

Figure 6. Parametric effect of depression on prefrontal arbitration control

(a) Depression effects on the model-free system's sensitivity to RPE (n=63; the total number of subjects, including 40 who participated behavioral experiment only and 23 who were also scanned with the fMRI). (Left) Relationship between the depression score and the learning rate parameter for reliability estimation of the model-free system. Individuals with a higher depression score tend to exhibit a higher learning rate, indicating that their reliability estimation for the model-free system is more sensitive to RPE. (Right) Illustrative examples of reliability changes of people with a low ("normal group") and a high depression score ("depressive group"), each of which is associated with a low and a high learning rate for reliability estimation, respectively. The depressive group shows rapid changes in MF reliability due to higher learning rate. (b) Relationship between the depression score and the estimated effect size of Max reliability for frontopolar cortex (FPC), the brain area implicated in arbitration control (n=28; the total number of subjects scanned with fMRI). The estimated effect size of the Max reliability signal for FPC (coordinates [8,44,40]; z-score=3.83) is negatively correlated with the depression score. (c) Comparison of reliability signal representation performance in normal (n=15) and depressive (n=13) group (GLM and MVPA analysis). The parameter estimates of the Max reliability for the bilateral ilPFC ([-52,26,16], z-score=4.48; [42,20,-8], z-score=3.61) and FPC ([8,44,40], z-score=3.83) tend to decrease in the depressive group. The MVPA analysis with these three seed regions reveals that the amount of information about the model-free reliability was significantly higher in the depressive group in all three regions (one-way ANOVA). Asterisk (*) indicates significant difference at the 0.05 level.

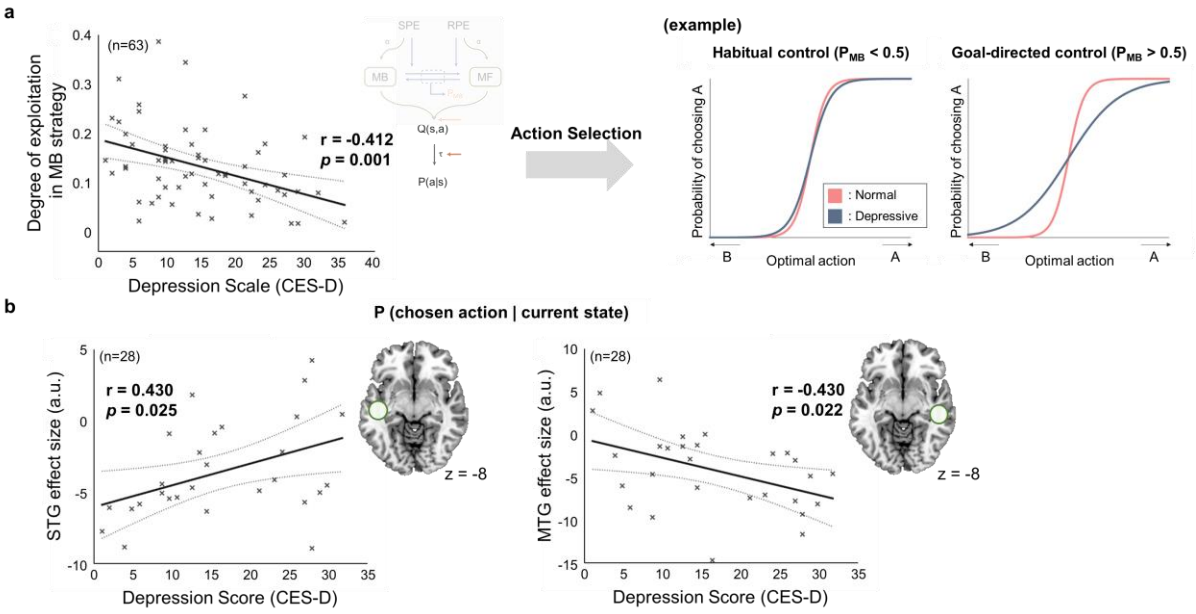

Figure 7. Parametric effect of depression on value-action conversion

(a) Depression effects on model parameters (n=63; the total number of subjects, including 40 who participated behavioral experiment only and 23 who were also scanned with the fMRI). (Left) Relationship between the depression score and the degree of exploitation in MB strategy (τ_{MB}). The MB exploitation parameter decreases as the depression score increases. (Right) Examples illustrating exploitation parameter effects on action selection. Shown are the softmax functions that convert an action value into a choice probability value, between the normal and the depressive group for goal-directed and habitual control. Compared to the normal group (pink), the depressive group makes more exploratory choices especially as they rely more on the model-based system (blue). (b) Relationship between the depression score and the parameter estimate of the probability of selecting chosen action for the two seed regions, right Superior Temporal Gyrus (STG) ([-46,-20,-8]; z-score=4.21) and the left Middle Temporal Gyrus (MTG) ([56,-32,-8]; z-score=3.55). The effect size of right STG and left MTG increases and decreases with the individual severity of depression, respectively.