Memory-guided goal-driven reinforcement learning explains subclinical depression

Gyubin Lee*
School of Computing
Korea Advanced Institute of Science and Technology
Republic of Korea, Daejeon 34141
gbstorm81@kaist.ac.kr

Minsu Abel Yang
Department of Bio and Brain Engineering
Program of Brain and Cognitive Engineering
Korea Advanced Institute of Science and Technology
Republic of Korea, Daejeon 34141
minsuyang@kaist.ac.kr

Sang Wan Lee†
Department of Bio and Brain Engineering
Program of Brain and Cognitive Engineering
Center for Neuroscience-inspired Artificial Intelligence
KAIST for Health Science Technology
KAIST for Artificial Intelligence
Korea Advanced Institute of Science and Technology
Republic of Korea, Daejeon 34141
sangwan@kaist.ac.kr

Abstract

Depression often leads to impairment in various cognitive functions. Despite earlier computational and neural studies, we have a limited understanding of how depression affects goal-driven and memory-guided value learning. First, we designed a computational model of reinforcement learning (RL) that incorporates memory-based control into the successor representation (SR), called goal-driven SR. Second, using behavioral and fMRI data acquired using a two-stage goal-directed decision-making task, we found evidence for goal-driven SR; SR prediction error was encoded in the neural activity of the ventrolateral prefrontal cortex (vlPFC) and right insula, whereas the conventional reward prediction error was encoded in the ventral striatum. The results suggest the potential role of the prefrontal cortex and insula in memory-based value learning. Third, in the model-based correlation analysis, we found the effect of depression on value learning, memory-based control, and action selection. Lastly, we built a novel Predictor-Decoder system that estimates the severity of depression from the goal-driven SR parameters, with significantly higher accuracy than data-driven machine learning models, including support vector regression (SVR) and deep neural network (DNN).

Keywords: Depression, Value-based decision-making,
Memory-based control, Prediction of depression severity

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2019M3E5D2A0106626721, Development of metacognitive AI for rapid learning), and Institute of Information & Communications Technology Planning& Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence).

*First author
†Corresponding author

1 Introduction

Depression is known to affect human value-based decision-making. While recent studies have attempted to understand depression in the context of reinforcement learning (RL), whether and how depression affects goal-driven and memory-guided learning remains largely unknown. Although the recent study found that arbitration between model-based and model-free systems can characterize depression well [1], the existing theory does not fully accommodate memory involvement in goal-driven learning. This necessitates the understanding of depression in the context of goal-directed and memory-guided value learning.

This study investigates how subclinical depression affects memory-guided goal-directed learning at the computational and neural levels. First, we implemented a computational model incorporating memory-based control into successor representation (SR). We ran a model comparison analysis on behavioral data to show that our model explains human subjects' choice patterns qualitatively better than model-based and model-free RL. Second, we conducted general linear model (GLM) analysis on fMRI data. Next, in the correlation analysis, we showed that the model parameters contained information about depression. Finally, we built a framework to predict the severity of depression based on the parameter of our model, demonstrating the possibilities of model-based diagnosis of depression.

2 Result

2.1 Computational model of memory-guided goal-directed learning

To establish computational principles of memory-guided goal-directed learning, we constructed the computational model based on SR named 'goal-driven SR' (Figure 2 left) and fitted its parameter to behavioral data while subjects were performing the two-stage Markov decision task (Figure 1). The task involves a goal condition, in which a subject needs to make a choice leading to a specific goal that is assigned on a trial-by-trial basis. We constructed a "memory bank" in the computational model [4] to account for the suboptimal learning and decision-making process caused by maladaptive memory functions originated from depression. The memory bank keeps weight vectors for each goal condition in the task and returns the weight vector associated with the given goal condition. With support of the memory bank, the model can explicitly learn a goal-specific policy, whereas a previous RL algorithms maintained one policy to multiple goals, which update it whenever the goal changes by running backward planning [2, 3].

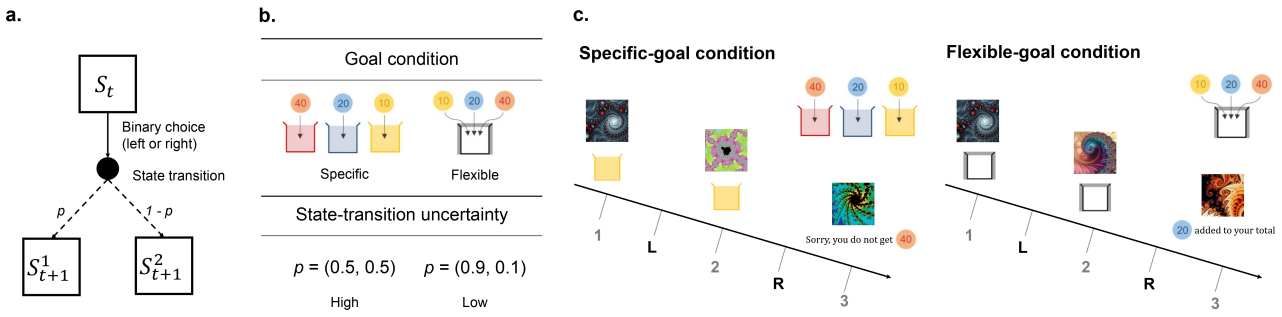


Figure 1: (a) Sequential two-stage Markov decision task proposed in [3]. Participants make binary choice (left or right) at each state and move to one of next two states with a certain state-transition probability (p, $1-p$) (b) Illustration of the task context. The goal condition consists of specific and flexible-goal conditions. In specific-goal condition, the color of the collecting box (red, blue or yellow) should match the color of the coin. In flexible-goal condition, all kinds of coins are allowed to be collected. The state-transition uncertainty consists of high and low uncertainty conditions. During state-transition uncertainty is high, state-transition probability was set to $p = 0.5$, whereas during the low uncertainty condition, state-transition probability was set to $p = 0.9$. (c) Illustration of the task. Each state is associated with a specific fractal image. After the fractal image for initial state was shown, participants made a binary choice (left or right). Then, they were transited to the next state determined by the state-transition probability. Again, subjects made a binary choice again to arrive to an outcome state. Each outcome state was associated with a colored coin (red, blue, yellow). In the specific-goal condition, participants should collect the coin that matches with the color of the collecting box. In the flexible-goal condition, they can receive the reward afterward regardless of the color of coin.

We compared the prediction performance of our computational model with that of the goal-directed RL model by utilizing Bayesian information criterion (BIC). As a result, we found that the BIC score goal-driven SR model (M=614.22, SEM=20.56) is significantly lower than the BIC score of arbitration control model (M=666.83, SEM=18.33) (Figure 2c-d, paired-sample permutation test, goal-driven SR model vs. arbitration control model: $p < 10^{-5}$, $d = -0.8524$, $n = 64$).

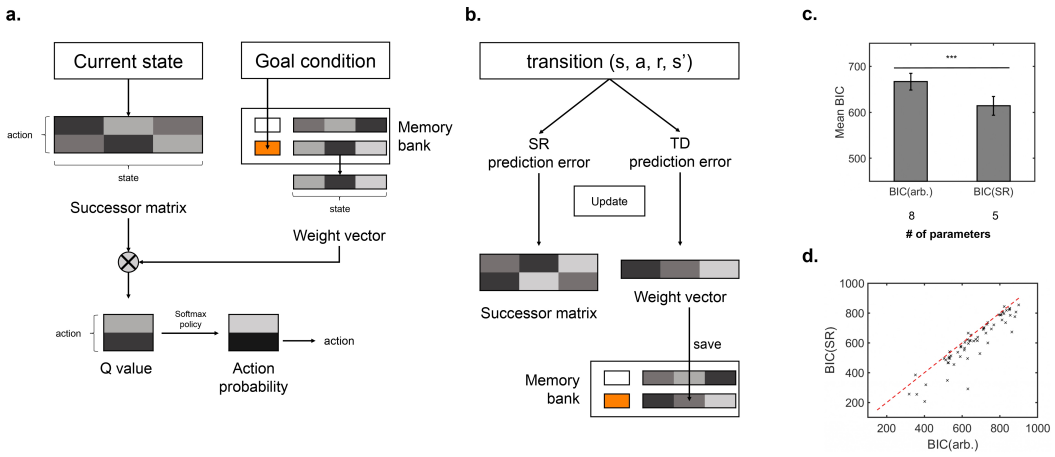


Figure 2: (a) The structure of goal-driven SR. In each state, the model returns a successor matrix that indicates expected cumulative future state occupancy for each action. When the new goal is assigned, the memory bank returns the saved weight vector associated with the goal condition. The state-action value (Q value) is given by the product of the successor matrix and the weight vector. Using the state-action value, the model makes a stochastic choice with the softmax function. (b) After transition (s, a, r, s'), two types of prediction error occur: SR prediction error and temporal difference (TD) prediction error. SR prediction error refers to the difference between actual and expected state occupancy, and TD prediction error refers to the difference between received and predicted reward. The model updates its successor matrix using SR prediction error and weight vector using TD prediction error. After updating the weight vector, the memory bank saves it, tagging the corresponding goal condition. (c-d) Model comparison analysis result. We used Bayesian information criterion (BIC); (c) The mean and standard error of the mean of BIC scores. Error bar stands for the standard error of the mean, (d) Individual comparison of BIC scores for each subject. X axis indicates BIC of arbitration control, and Y axis indicates BIC score of goal-driven SR. The dotted line in the figure indicates identity line as reference.

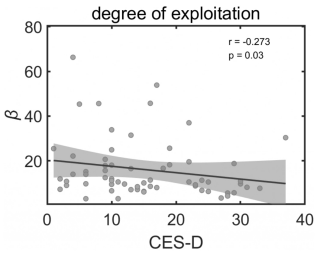
2.2 Neural evidence of goal-driven SR

To verify that our model explains the neural activity in the brain, we conducted general linear model (GLM) analysis to check whether the key signals from our model, SR and TD prediction errors, were represented within the fMRI data. We found that the SR prediction error was found in the right insula and the ventrolateral prefrontal cortex (vlPFC) (Figure 3 left, all $p < 0.05$, FWE corrected for multiple comparisons). On the other hand, the TD prediction error was found on both sides of ventral striatum (Figure 3 right, all $p < 0.05$, FWE corrected for multiple comparisons).

a.

Parameter	α_{TD}	α_{SR}	γ_{SR}	β	b_0
r	-0.1408	0.0509	-0.0389	-0.2729	-0.0327
p	0.2710	0.6920	0.7629	0.0304	0.7993

α_{TD}: learning rate of weight vector, α_{SR}: learning rate of successor matrix, γ_{SR}: discount factor of successor matrix, β: degree of exploitation, b_0: bias in action selection



b.

Parameter	α_{TD}	α_{SR}	γ_{SR}	β	b_0
r	-0.3598	-0.0315	-0.3967	-0.4828	-0.0114
p	0.0051	0.8127	0.0019	0.0001	0.9315

α_{TD}: learning rate of weight vector, α_{SR}: learning rate of successor matrix, γ_{SR}: discount factor of successor matrix, β: degree of exploitation, b_0: bias in action selection

Figure 3: GLM analysis results. The SR prediction error were encoded in the neural activity of ventrolateral prefrontal cortex (vlPFC) and right insula, whereas the conventional TD prediction error were encoded in ventral striatum. All the effects reported were survived after adjusting for family-wise error (FWE $p < 0.05$).

2.3 Model parameters explain depression severity

To verify that our model reflects the characteristics of the subclinical depression, we measured Spearman's correlation coefficients between each subject's estimated model parameter and their self-reported depression score (CES-D). First, we found that the degree of exploitation is negatively correlated ($r_S(62) = -0.273, p = 0.03, n = 63$) with CES-D score (Figure 4 left.) For more detailed investigation, we conducted partial correlation analyses and found that the learning rate of the weight vector ($r_S(62) = -0.360, p = 5.10 \times 10^{-3}, n = 63$), the discount factor of the successor matrix ($r_S(62) =$

$-0.397, p = 1.90\times10^{-3}, n = 63$), and the degree of exploitation ($r_S(62) = -0.483, p = 1.00\times10^{-4}, n = 63$) are significantly negatively correlated with CES-D score (Figure 4 right.)

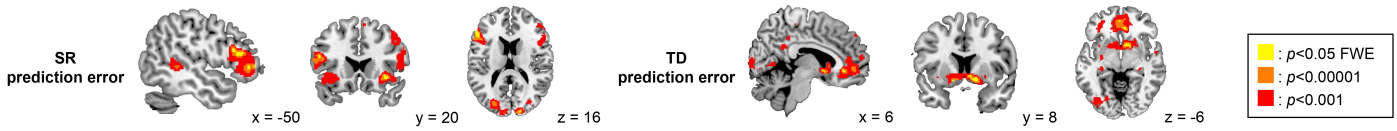


Figure 4: (a) Simple correlation analysis between model parameters and CES-D score. Right graph shows the relationship between the individual CES-D score and the degree of exploitation, one of the model parameters. (b) Partial correlation analysis between model parameters and CES-D score.

2.4 Model-driven diagnosis of depression

To fully accommodate the observation that our model parameters convey information related to depression, we designed a framework that predicts the subject's CES-D score using model parameters. First, we attempted to directly predict subjects' CES-D score from their fitted parameters using data-driven approaches, including support vector regression (SVR) and deep neural network (DNN). However, their performance was poor mainly because there were insufficient severely depressed subjects compared to the control case. To resolve this issue, we built a Predictor-Decoder system, which consists of the 'predictor' and 'decoder' modules (Figure 5a). The predictor receives the individually-fitted parameter set of goal-driven SR as input (Figure 5b). Then it compares the input parameter vector with the parameters of other subjects whose CES-D scores are known by the predictor. The outputs of the predictor are 1) a Boolean value set; each element indicates whether the subject's CES-D score is higher than that of the input subject, and 2) the posterior probability of the prediction. The decoder converts the output of the predictor into the subject's CES-D score by using the iterative hypothesis evaluation (Figure 5c). First, the decoder assumes the subject's CES-D score as an integer M that starts with 1 (Hypothesis H_M). Then it evaluates H_M by calculating the score of H_M. After evaluating all hypotheses, the decoder accepts the hypothesis with the highest score. Notably, our Predictor-Decoder system can predict the CES-D score well (Figure 5d).

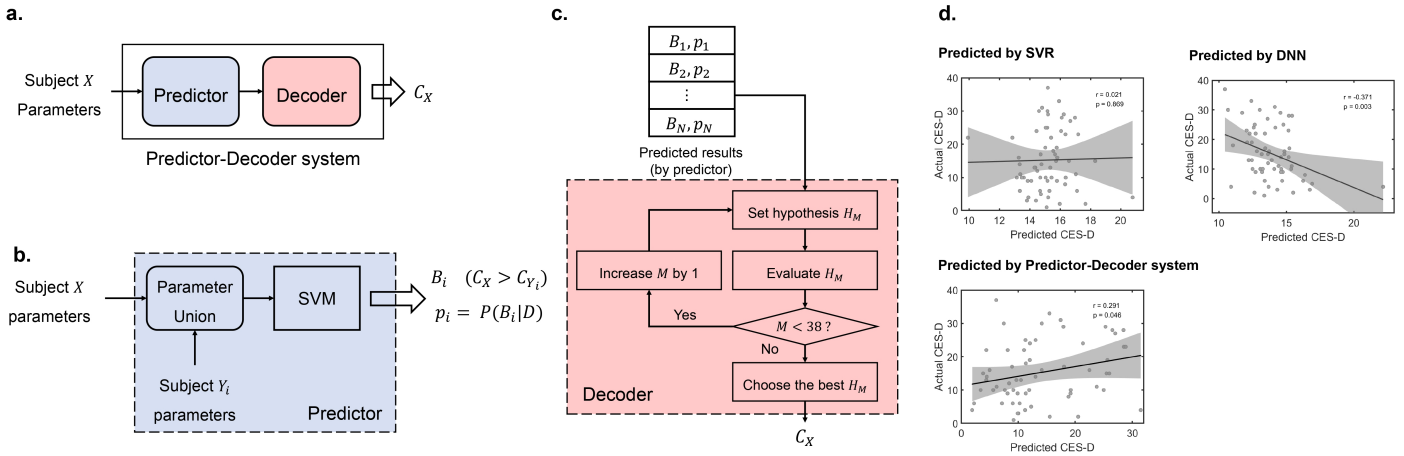


Figure 5: (a) The outline of the Predictor-Decoder system. The parameters of goal-driven SR for the subject X are fed into the system as an input. The model outputs the predicted CES-D score of subject X C_X. (b) The structure of the predictor. Model parameters of subject X are combined with model parameters of subject Y_i, whose CES-D score is known. The concatenated parameter vector is passed to the binary SVM classifier that outputs Boolean value B_i (True if C_X is higher than CES-D score of subject Y_i C_{Y_i}, False otherwise) and the posterior probability of the prediction $p_i = P(B_i|D)$. (c) Decoder is an algorithm that estimates a subject's CES-D score using the predictor's output. The decoder assumes C_X being in the range from 1 to 37 (= a maximum CES-D score in our dataset), each of which is considered to represent a different 'hypothesis'; the decoder evaluates each hypothesis, iteratively. The score of H_M, S_{H_M} is $S_{H_M} = \sum_{i=1}^{N} a_i \times p_i + (1 - a_i) \times (1 - p_i)$ (N is the number of predicted results by the predictor. a_i is 1 if B_i is a right prediction under H_M and 0 otherwise). (d) Prediction performance of support vector regressor (SVR), deep neural network (DNN), and our Predictor-Decoder system. The predicted CES-D score by SVR and DNN have no positive correlation with actual CES-D score, whereas the predicted score by our Predictor-Decoder system has significantly positive correlation with the actual CES-D score

3 Discussion

3.1 Neural evidence of goal-driven SR

Previous findings show the neural correlates of state prediction error (SPE) for the model-based learning in the insula and the lateral prefrontal cortex (lPFC), whereas association with reward prediction error (RPE) for the model-free learning appears in the ventral striatum [3]. We found that goal-driven SR is associated with neural representations of the two distinct prediction errors: SR and TD prediction error. The successor matrix keeps track of expected cumulative future state occupancies. In our task, the visited state cannot be re-visited in the same trial, so the successor matrix should be converged to the state-transition probability. Accordingly, the SR prediction error resembles SPE. In our analyses, SR prediction error was found in the insula and ventrolateral prefrontal cortex (vlPFC), consistent with previous findings of SPE. Taken together, the results expand the potential role of the lateral prefrontal cortex and insula to memory-guided goal-driven learning. The TD prediction error is associated with the ventral striatum like RPE of the model-free system. Although the updating processes are quite different, the concept of TD prediction error and RPE are similar, therefore it can be an evidence that the TD prediction error works like RPE.

3.2 Effects of depression on memory-guided goal-driven learning

For the first, the arbitration control [1] and our goal-driven SR have a parameter that indicates the tendency to make the most valuable choice. The corresponding parameter value is negatively correlated with the CES-D score in both the arbitration and our model. This result suggests that the depressed group makes more exploratory choices rather than exploiting the previous history of success than the healthy group. Also, we found that the learning rate of temporal difference learning is negatively correlated with the CES-D score. It indicates that the depression group takes more time to update state values than the healthy control group. In this case, if the subject happens to obtain more rewards than previously, the control group updates their state-action value accordingly, whereas the depressed group is less capable of improving their policy. This can be seen as habitual behavior. This result corroborates a previous finding that choice optimality is negatively correlated with CES-D score. One interesting result is that the discount factor for the successor matrix is negatively correlated with the CES-D score. The successor matrix provides information about the task structure. Therefore, this result shows that goal-driven SR can represent the effect of depression on memory-guided learning.

3.3 Predictor-Decoder system for predicting the depression score

We designed a Predictor-Decoder system to estimate CES-D score from model parameters, and it outperforms a few other data-driven models. The key idea is to learn the 'gradient' between two data points. The predictor is trained to compare subject's CES-D scores from the two subjects' model parameters. This strategy is intended to learn the difference in the CES-D score between two subjects instead of the CES-D score itself. Another advantage of the predictor design is that pairing subjects' model parameters can increase the number of training data. In addition, the decoder is intended to recover the subject's CES-D score correctly by obscuring the wrong predictions by the predictor. Although the most reasonable hypothesis may receive some fewer points ($(1 - P(B_i|D))$ because of the wrong prediction, the effects of incorrect prediction are counterbalanced after summing up the received points.

References

[1] Suyeon Heo, Yoondo Sung, and Sang Wan Lee. Effects of subclinical depression on prefrontal–striatal model-based and model-free learning. *PLoS computational biology*, 17(5):e1009003, 2021.

[2] Dongjae Kim, Jaeseung Jeong, and Sang Wan Lee. Prefrontal solution to the bias-variance tradeoff during reinforcement learning. *Cell reports*, 37(13):110185, 2021.

[3] Sang Wan Lee, Shinsuke Shimojo, and John P O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014.

[4] Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adria Puigdomenech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *International Conference on Machine Learning*, pages 2827–2836. PMLR, 2017.