2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Computational Simulation of Wisconsin Card Sorting Task by using Variational Recurrent Neural Network based on the free energy principle

Daiki Goto[1] and Hayato Idei[1,2] and Tetsuya Ogata[1,3]

***Abstract*— The Wisconsin Card Sorting Task (WCST) is used to measure cognitive flexibility. In WCST, the participants are required to estimate underlying rules (called "category" in WCST) from given sensory signals. Computational modeling of the underlying cognitive mechanisms of WCST is important for elucidating flexible cognitive processing. In this study, we propose a hierarchical Recurrent Neural Network (RNN) model for explaining the underlying cognitive mechanisms of WCST, based on the free energy principle (FEP). FEP explains perception and goal-directed action as the minimization of prediction errors between predicted and real sensory signals (called free-energy minimization) and is expected to be an integrated theory of the brain. The primary characteristic of our model is that it considers free energy at future time steps, enabling it to correctly answer WCST as a goal-directed behavior based on the FEP. The simulation experiment showed that the proposed model successfully estimated the underlying categories to be estimated in WCST and correctly answer WCST. This indicates that the proposed model may provide mechanistic insights into flexible cognitive processing from the perspective of FEP.**

I. INTRODUCTION

Cognitive flexibility is the ability to adapt to new tasks, rules, and changes in the environment and is considered a part of executive function (the ability to control one's behavior to achieve goals). Various neuropsychological methods have been devised to assess cognitive flexibility, one of which is the Wisconsin Card Sorting Task (WCST)[1].

The experimental procedure is outlined as follows: One trial of WCST consists of two-time steps. In the first time step, one stimulus card and four target cards, with features of three categories–color, shape, and number–are provided to the participants (Fig.1), and the experimenter determines one target category (color, shape, or number). In the second time step, the participants select one target card that has the same features as the stimulus card based on the category they estimated. As the participants do not know the category set by the experimenter, they have to infer it from the correct/incorrect feedback given after their answers. The experimenter returns correct/incorrect feedback regarding the target card selected by the participants. Correct feedback indicates that the estimated category is correct, and the participants should continue using that category. Incorrect feedback indicates that the estimated category does not match the category set by the experimenter and that the participants should therefore switch to inference about the target category.

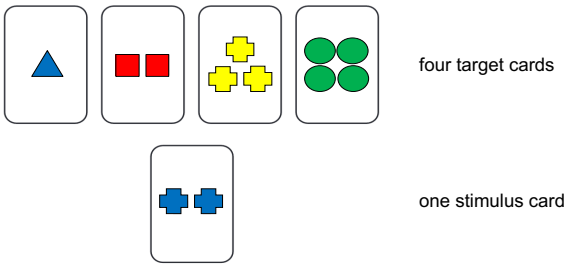

Fig. 1: Wisconsin Card Sorting Task. There are one stimulus card and four target cards, which have features about three categories: color, shape, and number. The participants select one card from the target cards which has same feature as stimulus card about one category set by the experimenter without sensory information of the category set by the exeperimenter.

This is repeated several times and the participants are expected to be able to appropriately estimate the category based on feedback and select the correct card. Furthermore, the target category is changed non-explicitly by the experimenter when the participants are able to answer correctly more than a certain number of times in a row. The participants should be aware of category changes based on feedback and be flexible in modifying their answer.

Computational modeling of the cognitive processes underlying the WCST is conducted to understand the underlying mechanisms. Several studies have been conducted by developing biologically valid reinforcement learning models modeling specific brain regions [2] [3] [4] [5]. As WCST performance decreases in several neurological and psychiatric disorders, these studies are expected to play an important role in clarifying relevant brain functions and mechanisms.

Recently, computational models of brain functions based on the free energy principle [6] have attracted considerable attention. Based on the framework of Bayesian inference, the free energy principle views biological brain functions as the output of predicted signals related to sensory signals, using prior beliefs about the external world and error minimization between predicted signals and sensory signals by

*This work was partially supported by JSPS Grantin-Aid (Nos. JP22J01708, JP22KJ3167) and JST Moonshot R&D (No. JPMJMS2031).

[1]Waseda University, 3-4-1 Ookubo, Shinjuku-ku, Tokyo, 169-8555, Japan https://www.waseda.jp/top/

[2]National Center of Neurology and Psychiatry, 4-1-1 Ogawa-Higashi, Kodaira, Tokyo, Japan https://www.ncnp.go.jp/hospital/

[3]National Institute of Advanced Industrial Science and Technology, 1-3-1 Kasumigaseki, Chiyoda-ku, Tokyo, Japan https://www.aist.go.jp

updating prior beliefs. The error to be minimized is called "free energy" (sometimes called "variational free energy"). Furthermore, the free energy principle theorizes that the generation of planning for goal-directed behavior is a form of free-energy minimization that differs from free-energy minimization(i.e., variational free energy minimization) in the perception of the world. In this framework, agents acquire goal-directed plans by predicting future sensory signals using prior beliefs about the external world and minimizing errors between predicted future sensory signals and desired future sensory signals (i.e., goals). As goal-directed planning generation can also be considered as a part of inference about sensory signals from the external world, the free energy principle refers to planning of goal-directed behavior generation and actions as "active inference" [7]. Thus, in goal-directed behavior, action plans are generated to minimize the free energy (sometimes called "expected free energy") within the framework of active inference.

While the free energy principle is expected to be a unifying theory of the brain because it can explain a wide range of cognitive phenomena, a computational approach to the WCST using the free energy principle has not been thoroughly explored. While previous studies have focused on local brain activity using a reinforcement learning model, examining the cognitive process of WCST within the framework of more general principles is expected to contribute to elucidating the relationship between the answering process of WCST and other cognitive functions.

In this study, we regard the answering process of WCST as goal-directed behavior (i.e., the participants are tasked with selecting the correct card) based on the framework of active inference. Based on this, we propose a hierarchical Recurrent Neural Network model (RNN) that can correctly answer the simulation task of WCST, which is constructed as a numerical simulation task by minimizing the free energy (expected free energy and variational free energy). Simulation experiments showed that the proposed model could appropriately estimate the category set by the experimenter and change its answers along with the change in the category. This suggests that the proposed model may provide insights into flexible cognitive processes in terms of the free energy principle.

II. Experiment

We set up a simulation experiment of WCST performed by the RNN model. In the standard WCST, the participants select one of four target cards with the same features in one category as the stimulus card. We regard these process as selecting the content of the estimated category of a stimulus card to simplify WCST for a simulation. The simulation experiment is outlined as follows.

Each trial of the simplified WCST experiment comprised two steps. In the first-time step, a stimulus card with features related to the three categories was presented. The target category was set by the experimenter. In the second time step, the RNN model selected features corresponding to the target category from the features of the three categories of stimulus cards. As the participants did not know the category set by the experimenter, they had to infer it from the correct/incorrect feedback given after their answers. The experimenter returned correct/incorrect feedback regarding the features selected by the participants. Correct feedback indicated that the estimated category is correct, and the participants should continue using that category. Incorrect feedback indicated that the estimated category does not match the category set by the experimenter and that the participants should switch to inference about the target category.

This was repeated 100 times, and the participants were expected to be able to appropriately estimate the category based on feedback and select the correct feature. Furthermore, the target category was changed non-explicitly by the experimenter at random time steps. The participants should be aware of category changes based on feedback and be flexible in modifying their answer. The experiment consists of two sessions: learning and test. Each session is outlined below.

A. Learning

During the learning session, we used 6-dimensional time series data as sensory data for the RNN model. The six dimensions consisted of stimulus card features (three dimensions), answer (one dimension), correct/incorrect signal of the answer (one dimension), and category set by the experimenter (one dimension).

For each category of the stimulus card, the features were assigned as 0.1, 0.3, 0.5, or 0.7. The values for the three categories differed. The answer was one of the three features of the stimulus card. The correct/incorrect feedback was expressed as 0.8 (in case of correct answer) or -0.8 (in case of incorrect answer). Information on the category set by the experimenter was expressed as -0.4, 0.0, or 0.4 (corresponding to three categories).

The definition of correct was the difference between the value (0.1, 0.3, 0.5 or 0.7) of the model's answer and its sensory signal was less than 0.2 Note that in accordance with the standard WCST, the category set by the experimenter was not given to the model in the test session.

One sequence of data consisted of 200 time steps (0, 1, 2,...). Each trial of WCST consisted of two steps: card recognition and answering. In the card recognition time step, a stimulus card was given, giving meaningful values only to the information on the three features of the stimulus card and the category set by the experimenter. For the information of the answer and the correct/incorrect signal, we gave 0.0 as a meaningless value (i.e., a value not used in the sequence). In contrast, at the answering time step we assigned meaningful values to all information because these were the answering steps. At these time steps, the three features of the stimulus cards were the same as those in the previous time step.

The RNN model was given 15 sequences of data and learned them. These consisted of three sequences of data in which the answer is always correct, six sequences in which the answer is always incorrect, and six sequences in

which the category set by the experimenter was changed once within the sequence and the answer was always correct.

B. Test

In the test experiment, we evaluated the ability of the RNN model to respond to a test data sequence that was not provided as learning data. The data were prepared in the same manner as the learning data; however, in accordance with the standard WCST, information on the category set by the experimenter was not provided to the model. Therefore, at even time steps in which the stimulus card is given, we assigned meaningful values only for the information of the three features of the stimulus card. During the test, the category set by the experimenter was changed without notice, and the RNN model was tasked with changing the estimation of the category used.

III. Proposed Model

We propose a computational model for the WCST based on the free energy principle using a predictive coding-inspired Variational RNN (PV-RNN). The PV-RNN was proposed by Ahmadi et al.[8] as an RNN to clarify the cognitive basis of the free energy principle. It is based on a multiple-timescale neural network (MT-RNN)[9]; therefore, the PV-RNN consists of deterministic neurons d. However, in addition to the deterministic neuron d, the PV-RNN consists of latent neurons z that represent beliefs in the brain regarding the causes of external sensations as Gaussian distributions. These variables have prior z_p and posterior z_q distributions, each representing the estimated causes before and after the sensory signals are observed. Using these latent neurons z, deterministic neurons d generate predicted signals.

Our model has a four-layer hierarchical structure, as shown in Fig.2. The model consists of three first layers into which the three feature values of the stimulus card are predicted/input; a second layer into which the sensory signal of one of three numbers (i.e., answer) is predicted/input; a third layer into which sensory signals of information of the correct/incorrect answer and the category to be used are predicted/input; and a fourth layer that monitors them. In this figure, $\hat{x}_{t,i}$ is the ith predicted signal at the time step t. The first to third predicted signals are 3 features of the stimulus card, the fourth predicted signal is the answer, the fifth predicted signal is the correct/incorrect signal of the answer, and the sixth is the category set by the experimenter.

The deterministic neurons of the first, second, and third layers are computed from the posterior distribution of the latent variable, inputs from higher levels, and recurrent inputs. The deterministic neuron of the fourth layer is computed from the posterior distribution of the latent variable and recurrent inputs. The internal state $h_{t,i}^{(s)}$ and deterministic neuron $d_{t,i}^{(s)}$ of the i element in the sequence s at time step $t(t \geq 1)$ are calculated as follows:

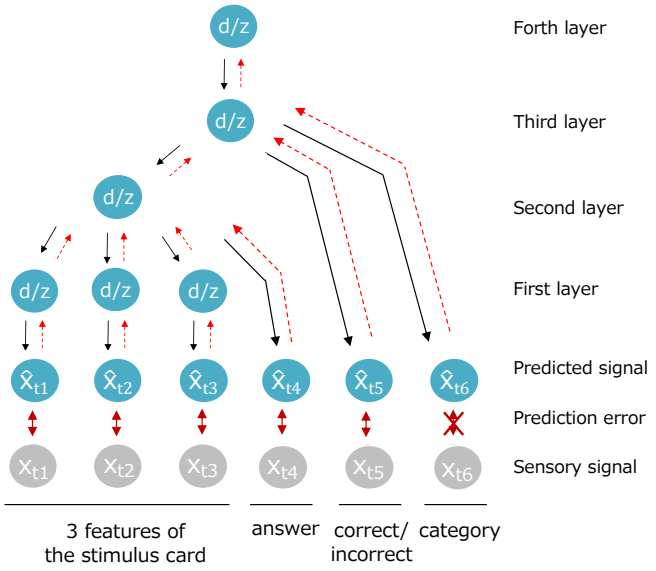

Fig. 2: Hierarchical PV-RNN. The model consists of three first layers into which the three feature values of the stimulus card are input, a second layer into which the sensory signal of one of three numbers (i.e. answer) is input, a third layer into which sensory signals of the correct/incorrect of the answer and the category set by the experimenter are input, and a fourth layer that monitors them.

$$
h_{t,i}^{(s)} = \begin{cases}
\frac{1}{\tau}\left(\sum_{j\in I_{\mathrm{Fod}}} w_{ij} d_{t-1,j}^{(s)} + \sum_{j\in I_{\mathrm{Foz}}} w_{ij} z_{t,j}^{(s)} + b_i\right) + \left(1-\frac{1}{\tau}\right) h_{t-1,i}^{(s)} & (i \in I_{\mathrm{Fod}}) \\
\frac{1}{\tau}\left(\sum_{j\in I_{\mathrm{Td}}} w_{ij} d_{t-1,j}^{(s)} + \sum_{j\in I_{\mathrm{Tz}}} w_{ij} z_{t,j}^{(s)} + \sum_{j\in I_{\mathrm{Fod}}} w_{ij} d_{t,j}^{(s)} + b_i\right) + \left(1-\frac{1}{\tau}\right) h_{t-1,i}^{(s)} & (i \in I_{\mathrm{Td}}) \\
\frac{1}{\tau}\left(\sum_{j\in I_{\mathrm{Sd}}} w_{ij} d_{t-1,j}^{(s)} + \sum_{j\in I_{\mathrm{Sz}}} w_{ij} z_{t,j}^{(s)} + \sum_{j\in I_{\mathrm{Td}}} w_{ij} d_{t,j}^{(s)} + b_i\right) + \left(1-\frac{1}{\tau}\right) h_{t-1,i}^{(s)} & (i \in I_{\mathrm{Sd}}) \\
\frac{1}{\tau}\left(\sum_{j\in I_{\mathrm{Fid}}} w_{ij} d_{t-1,j}^{(s)} + \sum_{j\in I_{\mathrm{Fiz}}} w_{ij} z_{t,j}^{(s)} + \sum_{j\in I_{\mathrm{Sd}}} w_{ij} d_{t,j}^{(s)} + b_i\right) + \left(1-\frac{1}{\tau}\right) h_{t-1,i}^{(s)} & (i \in I_{\mathrm{Fid}})
\end{cases} \tag{1}
$$

$$d_{t,i}^{(s)} = \tanh\left(h_{t,i}^{(s)}\right) \qquad \left(i \in I_{\mathrm{F_i d}}, I_{\mathrm{Sd}}, I_{\mathrm{Td}}, I_{\mathrm{F_o d}}\right). \tag{2}$$

Here, $I_{\mathrm{F_i d}}, I_{\mathrm{Sd}}, I_{\mathrm{Td}}, I_{\mathrm{F_o d}}$ represent the sets of indices of deterministic neurons in the first layer, second layer, third layer, and fourth layer, respectively, while $I_{\mathrm{F_i z}}, I_{\mathrm{Sz}}, I_{\mathrm{Tz}}, I_{\mathrm{F_o z}}$ represent the sets of indices of latent neurons in the first layer, second layer, third layer, and fourth layer, respectively. $w_{i,j}$ represents the synaptic weight from the jth neuron to the ith neuron. Additionally, $z_{t,j}$ represents the output of the posterior distribution at time step t for the jth neuron, τ represents the time constant, and b_i represents the bias in the ith neuron. The first layer contains three pairs of deterministic and latent neurons, each having synaptic connections only with the second layer and no synaptic connections within the first layer.

The latent neurons are represented as independent multivariate Gaussian distributions for each layer. The prior distribution is calculated from the previous deterministic neurons with mean μ and standard deviation σ as follows:

$$p\left(z_{t,i}^{(s)}\right) = p\left(z_{t,i}^{(s)} | d_{t-1,j}^{(s)}\right) = \mathcal{N}\left(z_{t,i}^{(s)}; \mu_{t,i}^{(s),p}, \sigma_{t,i}^{(s),p}\right). \tag{3}$$

$$\mu_{t,i}^{(s),p} = \tanh\left(\sum_j w_{i,j} d_{t-1,j}^{(s)}\right). \tag{4}$$

$$\sigma_{t,i}^{(s),p} = \exp\left(\sum_j w_{i,j} d_{t-1,j}^{(s)}\right). \tag{5}$$

The posterior distribution is calculated as follows:

$$\begin{aligned} q\left(z_{t,i}^{(s)} | e_{t:T^{(s)}}^{(s)}\right) &= \mathcal{N}\left(z_{t,i}^{(s)}; \mu_{t,i}^{(s),q}, \sigma_{t,i}^{(s),q}\right) \\ &\quad \left(i \in I_{\mathrm{F_i z}}, I_{\mathrm{Sz}}, I_{\mathrm{Tz}}, I_{\mathrm{F_o z}}\right), \end{aligned} \tag{6}$$

$$\mu_{t,i}^{(s),q} = \tanh\left(a_{t,i}^{(s),\mu}\right), \tag{7}$$

$$\sigma_{t,i}^{(s),q} = \exp\left(a_{t,i}^{(s),\sigma}\right), \tag{8}$$

$$z_{t,i}^{(s)} = \mu_{t,i}^{(s),q} + \sigma_{t,i}^{(s),q} \times \epsilon \tag{9}$$

Here, $T^{(s)}$ is the length of the data in sequence s. Also, ϵ is sampled from N(0,1). a is called the adaptation variable and is determined by the error e backpropagated by Backpropagation Through Time (BPTT). a is initialized by the initial internal state of the neuron that calculates the prior distribution before the learning or inference process.

$\hat{x}_{t,i}^{(s)}$ is the i the sensory prediction in sequence s at time step t. They are generated from deterministic neurons in the first layer, second layer, and third layer. The calculation of prediction generation is as follows. $I_{\mathrm{OF_i}}$, I_{OS}, and I_{OT} represent the set of neuron indices in the output layer connected to the first layer, second, third layer, respectively. The three preditions are generated by neurons in the first layer independently each other. In addition, the third layer generates two predictions.

$$\hat{x}_{t,i}^{(s)} = \begin{cases} \tanh\left(\sum_{j \in I_{\mathrm{F_i d}}} w_{i,j} d_{t,j}^{(s)}\right) & \left(i \in I_{\mathrm{OF_i}}\right), \\ \tanh\left(\sum_{j \in I_{\mathrm{Sd}}} w_{i,j} d_{t,j}^{(s)}\right) & \left(i \in I_{\mathrm{OS}}\right), \\ \tanh\left(\sum_{j \in I_{\mathrm{Td}}} w_{i,j} d_{t,j}^{(s)}\right) & \left(i \in I_{\mathrm{OT}}\right), \end{cases} \tag{10}$$

A. Parameter Update

The RNN model updates the synaptic weights and latent variables to minimize the variational free energy F calculated below.

$$\begin{aligned} F &= \sum_{s=0}^{S}\left(\sum_{t=0}^{T}\left(\sum_{i=1}^{8}\frac{1}{2}(x_{s,t,i} - \hat{x_{s,t,i}})^2\right.\right. \\ &+ \left.\left.\sum_{l=1}^{5} W^{(l)}\left(D_{KL}\left[q(z_t^{(l)}|e_{t:T})\middle\|p(z_t^{(l)}|d_{t-1}^{(l)})\right]\right)\right)\right) \end{aligned} \tag{11}$$

The first term is the prediction error term, and the second term is the KL divergence term. t is the time step, l is the layer index, s is the sequence index, and S is the number of sequences in the data. Additionally, d is the output of the deterministic neuron; z is the value of the latent variable; and p and q are the prior and approximate posterior distributions, respectively. Furthermore, e indicates the backpropagated error from t to the sequence end time step T. The parameter W is called the meta-prior. Ahmadi et al.[8] demonstrated the possibility of determining whether a model generates predictions deterministically or stochastically by adjusting this value.

B. Learning

Learning was performed by updating the RNN parameters (i.e., synaptic weights and adaptive variables) to minimize the variational free energy. The number of epochs was set to 50,000.

C. Test

By understanding the card selection process in WCST as goal-oriented behavior in terms of being tasked with selecting the correct card, we introduce the framework of active inference, which considered goal-oriented behavior to also be free energy minimization and set future time steps in the model in the test session to implement the framework. At a certain time step, the RNN model repeatedly generate predictions and update the parameters in the future, current, and past time steps by minimizing the sum of the free energies from the past time step to the future time step (Fig.3). Note that, in the test experiment, the sensory signals of the category were not given.

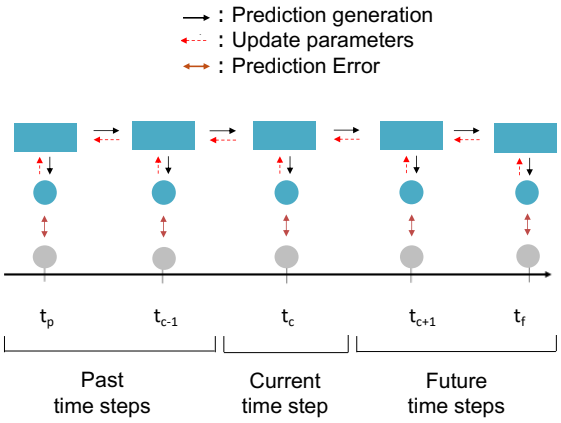

Fig. 3: In the test, at each time step, the model generates predictions and updates parameters at future time steps from t_{c+1} to t_f, current time step t and past time steps from t_p to t_{c-1}.

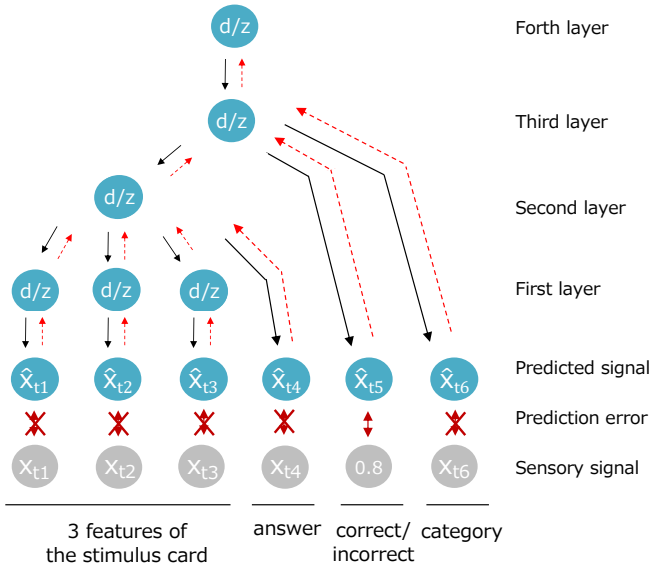

Fig. 4: Model at the future time steps. Only the prediction error is received from the correct/incorrect signal of the answer and takes the correct signal as the sensory signal, and then adjusts other outputs as well so that the free energy is minimized.

At the future time steps (Fig.4), after predicting all the sensory signals, only the prediction error was obtained from the correct/incorrect signal of the answer, taking the correct signal(i.e.,0.8) as the sensory signal. The other outputs of the model are then adjusted to minimize the free energy (i.e., the expected free energy G) at future time steps. This ensures that the model predicts the correct signal in the future and current time steps. A previous study[10] calculated the expected free energy in a PV-RNN, and based on this work, we introduced the expected free energy$G$.

The expected free energy G that is minimized at the future time step t_{c+t} $(> t_c)$ to generate the plan for goal-directed behavior is expressed in the following form: t_f indicates the furthest time within the range of future time steps to be considered. Further, e indicates the backpropagated error from t to t_f. The prediction error is calculated only for the correct or incorrect answer, and 0.8 (i.e., correct) is given as the sensory signal of the correct/incorrect answer.

$$G = \sum_{t=t_{c+1}}^{t_f} \left(\frac{1}{2}(0.8 - x_{\hat{t},i})^2 + \sum_{l=1}^{5} W^{(l)} \left(D_{KL}\left[q(z_t^{(l)}|e_{t:t_f}) \middle\| p(z_t^{(l)}|d_t^{(l)}) \right] \right) \right) \quad (12)$$

At the current time step (Fig.5), after outputting the predicted signals, the model receives the sensory signals of the three features of the stimulus card and the correct/incorrect signal of the answer. At this time step, no sensory signal for the answer was provided, in accordance with the actual WCST. Instead, the correct/incorrect signal is provided as a sensory signal of the correct/incorrect predicted signal of the answer output by the model at the current time step. The outputs of the model were adjusted to minimize the free energy (i.e., variational free energy F) calculated at this time step.

At the past time steps (Fig.6), after outputting a predicted signal, the model received sensory signals of the three features of the stimulus card, answer, and correct/incorrect answer. The sensory signals of the answer and correct/incorrect received at these time steps are the signals predicted by the model when this time step is the current time step. The outputs of the model were adjusted to minimize the free energy (i.e., the variational free energy, F) calculated at these time steps.

D. Parameter Setting

The number of deterministic neurons d in the first, second, third, and fourth layers was unified as 20, and the numbers of latent neurons z in the first, second, third, and fourth layers were 1, 3, 3, and 3, respectively. The time constant τ of the context layer, which was a parameter representing the temporal characteristics of a deterministic neuron d, was set to 1.0, 1.25, 1.5, and 4.0, for the first, second, third, and fourth layers, respectively. The value of the meta-prior W was set to 0.005 for all the layers. The parameters of the rectified Adam used for learning were as follows: learning rate = 0.001, $\beta 1$ = 0.9, and $\beta 2$ = 0.999. In the test, the range of past time steps was set to 3, 5, or 10 and the range of future time steps was set to 3. The number of epochs, which indicates the number of parameter updates at each time step, was set to 200, 300, or 400. The parameters of the rectified Adam were as follows: learning rate = 0.25, 0.5, or 0.75. The performance of the model was evaluated for all the combinations in a single test.

IV. Result and Discussion

To evaluate the model's performance, we recorded the correct answer, category prediction and numerical recognition

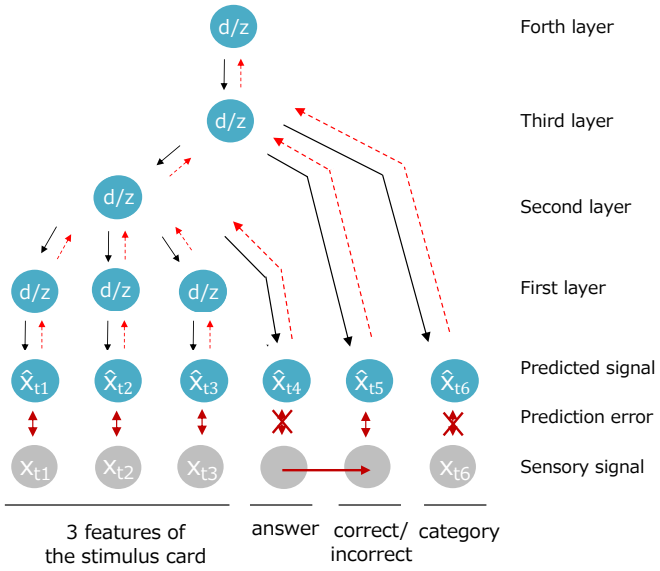

Fig. 5: Model at the current time step. The model receives sensory signals of three features of the stimulus card and the correct/incorrect signal of the answer. The sensory signal of the answer is not given to the model. The correct/incorrect signal is given as a sensory signal of the correct/incorrect of the predicted signal of the answer at the current time step.

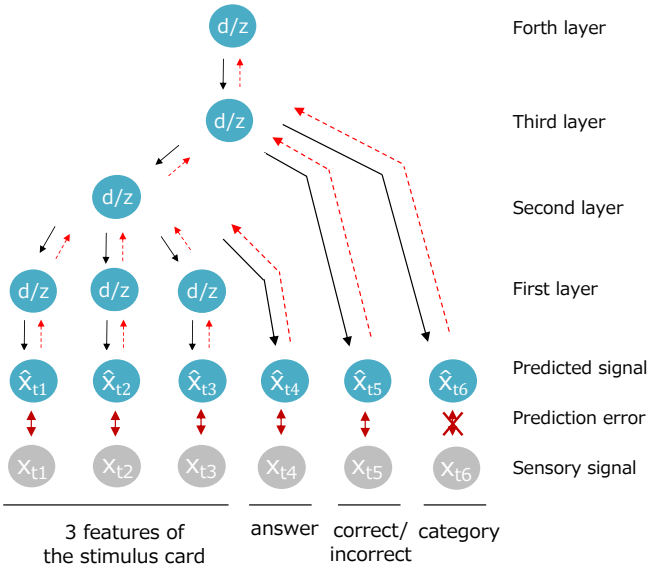

Fig. 6: Model at the past time steps. The model receives sensory signals of three features of the stimulus card, answer, and correct/incorrect of the answer.

category change	correct answer (%)	category prediction (%)	numerical recognition (%)
1 time	93	95	95
2 times	81	93	96
3 times	81	94	97

I. The result of the test experiment. The category set by experimenter is changed 1 time, 2 times, and 3 times. For each number of switching, tests were performed on 10 sequences of data. For each sequence, we tested all combinations of parameters (i.e., the range of past time steps, the range of future time steps, and the number of epochs) and recorded the result of all the combination with the highest correct answer rate as the test result of the sequence. This table shows the average of the results for 6 of 10 sequences, excluding the top two and bottom 2 sequences in the correct answer rate for each number of switches.

rates. The correct answer rate was defined as the probability that the difference between the value (0.1, 0.3, 0.5, or 0.7) of the answer and its sensory signal was less than 0.1, indicating that the model was able to answer the experiment correctly. The category prediction rate was defined as the probability that the difference between the estimated category and the category set by the experimenter was less than 0.2; it evaluates the model's ability to estimate the category. The numerical recognition rate was defined as the probability that the difference between the predicted signal of the three features of the stimulus card and the sensory signal was less than 0.1, which indicates that the model can properly recognize the given stimulus card.

The experiments were conducted by changing the category set by the experimenter once, twice, and thrice. For each switching number, tests were performed on 10 data sequences. For each sequence, we tested all combinations of the parameters (i.e., the range of past time steps, range of future time steps, and number of epochs) and recorded the score of the combination with the highest correct answer rate as the test result of the sequence.

Table 1 shows the average results for six of the 10 sequences, excluding the top two and bottom two sequences, in the correct answer rate for each number of switches. The correct answer rate was over 75%, the category prediction correct answer rate was over 80%, and the numerical recognition rate was over 90%.

Fig.7 show the temporal changes in sensory signals and predicted signals of the three features of the stimulus card, answer, correct/incorrect, and the category which changed once. In this figure, the blue line represents the sensory signals, and the red line represents the predicted signals. And Fig.8 shows the temporal changes in the mean and standard deviation of the posterior distribution of each layer.

Fig.7 shows that the value of the sensory signal (i.e., the blue line) of the category changes at time step 102. The category set by the experimenter was 0.4 until time step 102, and -0.4 after time step 102. The fact that the predicted signal (i.e., the red line) changed with changes in the sensory signal suggests that the RNN model can appropriately estimate the category set by the experimenter.

In Fig.8, the values of the mean(mu) and standard deviation(sigma) in the fourth layer change at time step 102 (i.e., the time step at which the sensory signal of the category

was changed by the experimenter). Note that the standard deviation temporarily decreases after the change in category. In addition, the posterior distribution of the first layer does not change significantly.

We held correlation analysis (cross-sectional analysis) on the result. The correlation coefficient between the mean posterior distribution of the first latent neuron in the second layer and the predicted signal for the first feature of the stimulus card was -0.661. Similarly, the correlation coefficient between the mean of the posterior distribution of the second latent neuron in the second layer and the predicted signal for the second feature of the stimulus card was 0.457 and that between the mean of the posterior distribution of the third latent neuron in the second layer and the predicted signal for the third feature of the stimulus card was 0.644. This suggests that the latent neurons in the second layer may reflect information regarding the three features of the stimulus card.

Similarly, for the posterior distribution of the third layer, we performed a correlation analysis of the time-series data (cross-sectional analysis) with respect to the predicted correct/incorrect signals. The correlation coefficient between the mean posterior distribution of the first latent neuron in the third layer and the predictive signal of correct or incorrect was 0.578. The correlation coefficient between the mean of the second latent neuron in the third layer and the predictive signal of the correct/incorrect is -0.807. The correlation coefficient between the mean of the third latent neuron in the third layer and the predictive signal of the correct/incorrect is -0.998. This suggests that latent neurons in the third layer may reflect information regarding correct or incorrect signals.

From the comparison with the sensory/predicted signals of the correct/incorrect and the category, it can be confirmed that the change in the mean and the standard deviation of the posterior distribution of the fourth layer occurs when an incorrect signal is given, and they do not change in answer to all incorrect signals but change only in answer to incorrect signals caused by the change in the category set by the experimenter.

An interpretation about the mechanism of the answering process of the proposed model is explained below. We show the flow of information of the model at the card recognition time step and at the answering time steps.Since the proposed model also have future time steps, we explain the flow of information when each time step is the future time step and when each time step is the current time step.

First, we explain the flow of information when recognition and answering time steps are at future time step. In these time step, the sensory signal of the correct/incorrect of the answer is given to the model as the correct signal. The prediction error signal between the sensory and predicted signal flows into the third layer, and the estimation of the latent variable in the third layer change. Since the third layer is involved in the calculation of all prediction signals, from the three features of the stimulus card to the category, when the estimation of the latent variables in the third layer change, all predicted signals change.

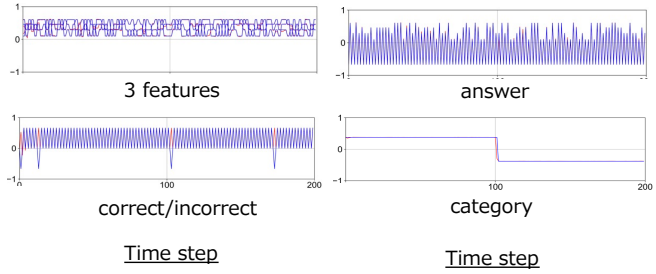

Fig. 7: Temporal changes in sensory signals and predicted signals of the 3 features of the stimulus card, answer, correct/incorrect, and the category in which the category set by the experimenter is changed once. The blue line shows sensory signals, and the red line shows predicted signals.

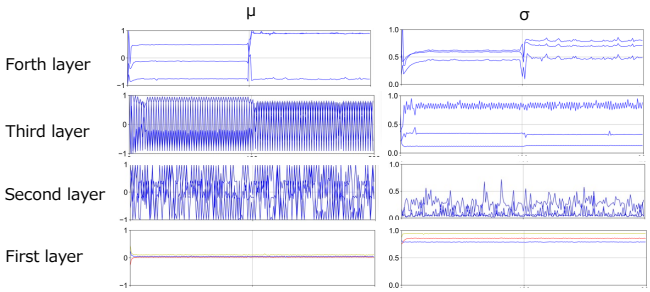

Fig. 8: Temporal changes in the mean(μ) and the standard deviation(σ) of the posterior distribution of each layer

Second, the flow of information when the card recognition time step is the current time step (Fig.9). At this time step, the model is randomly given sensory signals of the three features of the stimulus card. Because of the randomness, a prediction error occurs in most cases. The signal of the prediction error flows through the first layer to the second layer. As a result, the estimation of the latent variables in the second layer change. The change in estimation means that a given card is recognized by the model.

Third, we describe the flow of information when the answering time step is the current time step(Fig.10). In this case, the card recognition time step is the past time step. In this time step, if the predicted signal of the answer matches the sensory signal of it(i.e. the answer is correct), then the correct answer signal is given as the sensory signal of the correct/incorrect of the answer. As a result of setting the future time step, the model does not receive a prediction error signal of the correct/incorrect answer in this case, because the model predicts to receive the signal of the correct. Therefore, the estimates of the latent variables in the third layer and the forth layer are unchanged.

If the predicted signal of the answer does not match the sensory signal (i.e. the answer is incorrect), then the incorrect signal is given as the sensory signal. As a result, the network receives a signal of prediction error of the correct/incorrect signal. The signal of the prediction error flows to the third layer and the forth layer. Given the signal of prediction

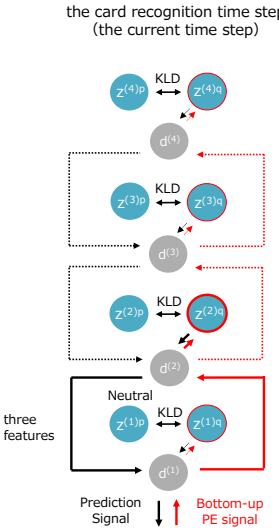

Fig. 9: The flow of information when the card recognition time step is the current time step. The signal of the prediction error caused by the sensory signal of the stimulus card flows through the first layer to the second layer. Then, the estimation of the latent variables in the second layer change.

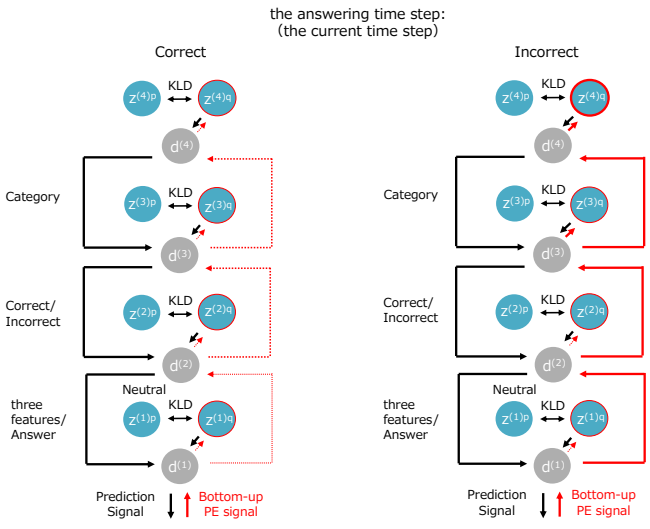

Fig. 10: The flow of information when the answering time step is the current time step. The signal of the prediction error caused by the sensory signal of the correct/incorrect flows to the third layer and the forth layer. Then, the estimation of the latent variables in the third layer and the forth layer change.

error, the estimation of the latent variable in the third and fourth layer are changed. The change in the estimation of the latent variable in the third layer results in a change in the prediction of the correct/incorrect signal. The change in the estimation of the latent variable in the forth layer results in a change in the prediction about the signal of the category. And predictions at subsequent time steps change accordingly. In this way, a model that receives an incorrect sensory signal changes its estimation of the category and changes its answer.

The model thus answers the experiment by changing the estimation of the latent variable corresponding to each signal so as to minimize the prediction error signal.

V. Conclusion

In this study, we proposed a hierarchical Recurrent Neural Network (RNN) model for explaining the underlying cognitive mechanisms of WCST, based on the free energy principle (FEP). These results are expected to contribute to the understanding of brain mechanisms related to WCST from the perspective of the free energy principle. In the future, it is expected that its performance will be compared with other computational models [2] [3] [4] [5]. Additionally, computational models using real images as stimuli are expected to be proposed.

References

[1] Berg, E,A.: A simple objective technique for measuring flexibility in thinking. Journal of General Psychology**39**(1), 15–22 (1948)

[2] Steinke, A., Lange, F., Seer, C., Kopp, B.:Toward a computational cognitive neuropsychology of Wisconsin card sorts: a showcase study in Parkinson's disease. Computational Brain and Behavior**1**(2), 137–150 (2018)

[3] Caso, A., Cooper, R.P.: A neurally plausible schema-theoretic approach to modelling cognitive dysfunction and neurophysiological markers in Parkinson's disease. Neuropsychologia**140**, 107359 (2020)

[4] Steinke, A., Kopp, B.: Toward a Computational Neuropsychology of Cognitive Flexibility. Brain Sciences**10**(12), 1000(2020)

[5] Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., Busemeyer, J. R.:Sequential learning models for the Wisconsin card sort task: assessing processes in substance-dependent individuals. Journal of Mathematical Psychology**54**(1), 5–13 (2010)

[6] Friston, K.: The free-energy principle a unified brain theory?. Nature Reviews Neuro-science**11**(2), 127–138 (2010)

[7] Friston, K., Samothrakis, S., Montague, R.: Active inference and agency: Optimal control without cost functions. Biological Cybernetics**106**, 523–525, (2012)

[8] Ahmadi, A., and Tani, J.: A novel predictive-coding inspired variational RNN model for online prediction and recognition. Neural Computation**31**(11), 2025–2074 (2019)

[9] Yamashita Y and Tani J.: Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. PLOS Computational Biology**4**(11), 1–18 (2008)

[10] Matsumoto, T., Ohata, W., Benureau, F.C.Y., Tani, J.: Goal-Directed Planning and Goal Understanding by Extended Active Inference: Evaluation through Simulated and Physical Robot Experiments. Entropy**22**(5), 564 (2020)