Decoding reward–curiosity conflict in decision-making from irrational behaviors

Yuki Konaka[1] and Honda Naoki[1, 2, 3, 4]*

1. Laboratory of Data-driven Biology, Graduate School of Integrated Sciences for Life, Hiroshima University, Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan
2. Kansei-Brain Informatics Group, Center for Brain, Mind and Kansei Sciences Research (BMK Center), Hiroshima University, Kasumi, Minami-ku, Hiroshima 734-8551, Japan
3. Theoretical Biology Research Group, Exploratory Research Center on Life and Living Systems (ExCELLS), National Institutes of Natural Sciences, Okazaki, Aichi, Japan.
4. Laboratory of Theoretical Biology, Graduate School of Biostudies, Kyoto University, Yoshidakonoecho, Sakyo, Kyoto 606-8315, Japan

***Corresponding author**: Honda Naoki
Address: Graduate School of Integrated Sciences for Life, 1-3-1 Kagamiyama, Higashi-Hiroshima City, Hiroshima 739-8526, Japan
Tel.: +81-82-424-7336
E-mail: nhonda@hiroshima-u.ac.jp

Keywords: Mental conflict; Curiosity, Reinforcement learning; Reward; Machine learning; State-space model.

Abstract

Humans and animals are not always rational. They not only rationally exploit rewards but also explore an environment, even if reward is less expected, owing to their curiosity. However, the mechanism of such curiosity-driven irrational behavior is largely unknown. Here, we developed a novel decision-making model for a two-choice task based on the free energy principle, which is a theory integrating recognition and action selection. The model successfully described irrational behaviors depending on the curiosity level. We then proposed a machine learning method to decode temporal curiosity from behavioral data, which enables us to quantitatively compare estimated curiosity and neural activities. By applying it to rat behavioral data, we found that the irrational choices sticking to one option was reflected to the negative curiosity level. Our decoding approach can be a fundamental tool for identifying the neural basis for reward-curiosity conflicts. Specifically, it could be effective in diagnosing mental disorders.

Introduction

Animals and humans perceive the external world through their sensory systems and make decisions accordingly[1,2]. Generally, they cannot make optimal decisions because of the uncertainty of the environment as well as the limited computational capacity of the brain and time constraints in decision-making[3]. In fact, they perform irrational actions. For example, people play lotteries and gamble despite low reward expectations. In this case, they face a dilemma between low expected reward and curiosity regarding whether a reward will be acquired. Thus, understanding how animals control the balance between reward and curiosity is important for the elucidation of the whole decision-making process. However, no method to quantify the reward–curiosity balance has yet been established. In this study, we developed a machine learning method to decode the time series of the reward–curiosity balance from animal behavioral data.

Some irrational behaviors emerge because of the strength of curiosity[4,5]. For example, conservative individuals avoid uncertainty and prefer to select an action that leads to predictable outcomes. Conversely, inquisitive individuals strongly desire to know the environment rather than rewards and prefer to select an action that leads to unpredictable outcomes. Rational individuals fall midway between these two extremes; in an ambiguous environment, they select an action to efficiently understand the environment, and if the environment becomes clear, they select an action to efficiently exploit the rewards. Thus, curiosity has a significant impact on behavioral patterns, and it is naturally thought that animals control the balance between reward and curiosity in a context-dependent manner.

Adaptive learning behaviors have been modeled primarily by reinforcement learning (RL), which is a theory for describing reward-seeking behaviors[6]. However, animals not only exploit rewards but also explore the environment, even without rewards, to minimize the uncertainty of the environment owing to their curiosity. Recently, the free energy principle (FEP) was proposed by Karl Friston under the Bayesian brain hypothesis, in which the brain optimally recognizes the outside world according to Bayesian estimation[7–9]. FEP addresses not only the recognition of the external world but also action selection, which minimizes the uncertainty of the recognition of the external world. Furthermore, FEP was integrated with RL, and then action selection was formulated by maximizing both reward and curiosity. Note that curiosity can be regarded as information gain, that is, the extent to which we expect our recognition to be updated by new observation through the action[10,11]. However, FEP assumes that the weighting of rewards and curiosity is always constant and thus cannot treat actual animal behaviors in which the weights of rewards and curiosity are expected to change over time. Hence, conventional theories such as RL and FEP are limited in describing the conflict between reward and curiosity.

The identification of the temporal variability of curiosity is important for elucidating the neural mechanisms of the reward and curiosity conflicts in decision-making. Many neuroscience studies have extensively examined the neural mechanisms of decision-making in the context of

RL[12–16]. RL has worked as a guiding principle in the quantitative understanding of reward-seeking behavior. However, RL does not focus on curiosity-driven behavior; therefore, it is difficult to quantitatively understand the reward–curiosity conflict. To overcome this difficulty, a method to decode the temporal variability of curiosity from behavioral data is essential. Such a method would enable us to analyze neural correlates with the temporal variability of curiosity and, consequently, clarify how the brain controls the balance of reward and curiosity in a context-dependent manner.

In this study, we extended FEP by incorporating a meta-parameter that controls the conflict dynamics between reward and curiosity, called the reward–curiosity decision-making (ReCU) model. The ReCU model is able to exhibit various behavioral patterns, such as greedy behavior toward reward, information-seeking behaviors with high curiosity, and conservative behaviors avoiding uncertainty. Moreover, we developed a machine learning method, called the inverse FEP (iFEP) method, to estimate the internal variables of decision-making information processing. By applying the iFEP method to behavioral time series in a two-choice task, we successfully estimated the internal variables, such as variations in curiosity, recognition of reward availability, and its confidence.

Results

ReCU model: decision-making with reward–curiosity dilemma in a two-choice task

Animals perceive the environment by inferring causes, such as reward availability from observation, and then make decisions based on their own inferences (**Fig. 1A**). In this study, we developed an ReCU model of a decision-making agent facing a dilemma between reward and curiosity in the case of a two-choice task, in which the agent selects either of two choices associated with the same rewards but with different reward probabilities. If the agent aims to maximize cumulative rewards, the agent must select an option with a higher reward probability. However, in animal behavioral experiments, even after they learned which option was associated with a reward, they did not exclusively always select the best choice, but also often selected the option with a smaller reward probability, which seems unreasonable. Consequently, we hypothesized the following: Animals assume that the reward probability for each option might fluctuate over time so that the continuous selection of one option decreases the confidence of the reward probability estimation for the other option. Thus, they become curious about the ambiguous option even with a smaller reward probability, and then selecting the ambiguous option is reasonable to increase the confidence of the estimation for both options. Therefore, we considered that the agent should make a decision driven by reward and curiosity in a situation-dependent manner.

In this two-choice task, the reward is given as all-or-none, but its intensity depends on the agent's own feeling, as

$$r = 0 \ or \ \log\left\{\frac{P_o}{1 - P_o}\right\},$$

where P_o is a parameter that controls the reward intensity felt by the agent. According to the Friston formulation, P_o represents the desired level of how much the agent wants the reward[10,11]. To model decision-making driven by reward and curiosity, we divided information processing in the brain into two processes. First, the agent recognizes the world, that is, it estimates the reward probability of each option (**Fig. 1A**; Process 1). Then, based on this recognition, the agent selects an action based on reward and curiosity (**Fig. 1A**; Process 2).

In Process 1, we modeled the recognition process of reward probability using sequential Bayesian updating. The agent assumed that the reward is probabilistically generated from the latent cause w for both options (Fig. **1B**, left), and these probabilities are represented by a sigmoidal function of w_i as

$$\lambda_i = f(w_i),$$

where λ_i and w_i denote the recognition of the reward probability of option i and its control variable, respectively, and $f(x) = 1/(1 + e^{-x})$ (Fig. **1B**, middle). In other words, either reward or non-reward was observed at the probabilities λ_i and $1 - \lambda_i$ if the agent selected option i.

In addition, the agent assumed an environment in which the reward probabilities for both options could change ambiguously over time (**Fig. 1C, upper panel**). To describe this, w is fluctuated by a random walk as

$$w_{i,t} = w_{i,t-1} + \sigma\xi_{i,t},$$

where $t, \xi_{i,t}$, and σ denote the trial of the two-choice task, Gaussian noise with zero mean and unit variance, and the noise intensity, respectively. Accordingly, the reward probabilities for both options can be described by $\lambda_{i,t} = f(w_{i,t})$. The temporal dynamics of the latent variable $w_{i,t}$ and reward observation are described by the state-space model from the viewpoint of the agent (agent-SSM) (**Fig. 1C, middle**).

Given the above environment, the agent aims to estimate the reward probability $\lambda_{i,t}$ of each option at each observation. To describe the confidence of $\lambda_{i,t}$, the agent is assumed to estimate the control variable $w_{i,t}$ as a distribution $P(w_{i,t})$ and update its distribution. In general, $P(w_{i,t})$ can be updated by action and reward observation o_t (**Fig. 1C, lower**), which can be roughly expressed as follows:

$$\text{updated belief } P(w_{i,t}) = \text{ previous belief } P(w_{i,t-1}) \times o_t.$$

This process was formulated using Bayes' rule (see Methods). We then derived the updates of the parameters as follows:

$$\mu_{i,t} = \begin{cases} \mu_{i,t-1} & (option\ i\ is\ not\ selected) \\ \mu_{i,t-1} - \alpha\left(\frac{1}{p_{i,t-1}} + \frac{1}{p_w}\right)\left(o_t - f(\mu_{i,t-1})\right) & (option\ i\ is\ selected) \end{cases}, \quad (1)$$

$$p_{i,t} = \begin{cases} \dfrac{p_{i,t-1}p_w}{p_{i,t-1}+p_w} & (option\ i\ is\ not\ selected) \\ \dfrac{p_{i,t-1}p_w}{p_{i,t-1}+p_w} + f(\mu_{i,t})(1-f(\mu_{i,t})) & (option\ i\ is\ selected) \end{cases}, \quad (2)$$

where $p_w = 1/\sigma^2$; $\mu_{i,t}$ and $p_{i,t}$ denote the mean and precision (i.e., inverse of variance) of $P(w_{i,t})$ (see the details in Methods), respectively. If the i-th choice is not selected, its precision decreases (i.e., $p_{i,t+1} < p_{i,t}$). With observation, belief is updated by the prediction error (i.e., $o_t - f(\mu_{i,t})$), and its precision is improved. Because the reward probability is represented by $f(w_{i,t})$, the estimated reward probability, that is, the recognized reward probability, should be transformed as $f(\mu_{i,t})$. Similarly, the confidence of the recognized reward probability should be evaluated not in $w_{i,t}$-space, but in $\lambda_{i,t}$-space, and hence the confidence is defined by $\gamma_t = p_{i,t}/f'(\mu_{i,t})^2$ (see Methods).

In Process 2, the agent selects actions according to the recognition of the reward probability and its confidence in Process 1. Agents have two kinds of motivations in decision-making: the desire to maximize the reward, and curiosity—the desire to update their beliefs with more confidence (**Fig. 1D**, **upper**). This sum, called "expected utility" in this study, can be expressed as

$$U(a_{t+1}) = c_t \cdot E[Information\ gain_{t+1}] + E[Reward_{t+1}],$$

where "information gain" represents the information derived from a new observation based on previous recognition, $E[x]$ denotes the expectation value of x based on current recognition, and c_t denotes a meta-parameter describing the intensity of curiosity, which weights the expected information gain. We regard c_t as varying over time (**Fig. 1D, lower panel**). We analytically derived $U(a_{t+1})$ as a function of $\mu_{t,a_{t+1}}$ and $p_{t,a_{t+1}}$ (see Methods). In decision-making, the agents selected action a_{t+1} with the highest expected utility with a higher probability as

$$P(a_{t+1}) = \frac{exp\big(U(a_{t+1})\big)}{\sum_a exp\big(U(a)\big)}, \quad (3)$$

which is a SoftMax function.

Sequential recognition and decision-making in the simulation

To validate our model, we performed simulations for two cases. In the first case, where reward probabilities were constant and different between the two options (**Fig. 2A**), the agent preferred to select the option with the higher reward probability (**Fig. 2B**). The recognized reward probabilities converged to ground truths, indicating that the agent accurately recognized the reward probabilities (**Fig. 2C**). The recognition confidence changed over time depending on the behavior; the confidence increased for the option that was selected more (**Fig. 2D**). The expected information gain was lower for the option with higher confidence (**Fig. 2E**), whereas the expected reward consistently followed the recognized reward probability (**Fig. 2F**). The expected utility, which is the sum of the expected information gain and reward, represents the value of each selection (**Fig. 2G**), resulting in the agent preferentially selecting the option with the higher expected utility.

In the second case, we assumed a time-dependent reward probability (**Fig. 2A'**). In the simulation, the agent adaptively changed its recognition of the reward probability following the change in the true reward probability and selected the option with the higher estimated reward probability (**Fig. 2B' and C'**). The confidence was also affected by the uncertainty of the reward probability at each time; the confidence was high where the reward probabilities were near-deterministic, around one and zero, and low where the reward probabilities were uncertain, approximately 0.5 (**Fig. 2D'**). The expected information gain of an option was negatively correlated with the confidence of the option (red line in **Fig. 2E'**), suggesting that the agent was curious about the uncertain option. As in the above case, the expected reward varied depending on the recognized reward probability (**Fig. 2F'**). The expected utility changed similarly to the expected reward, but the difference between the left and right options was less pronounced owing to curiosity (**Fig. 2G'**). These two demonstrations indicate that our model is able to represent the process of cognition and decision-making based on reward and curiosity.

Curiosity-dependent irrational behaviors

Next, we examined how behavioral patterns are regulated by the intensity of curiosity and the degree of reward-seeking (**Fig. 3**). In a situation where the reward probabilities were zero for the left and 0.5 for the right (**Fig. 3A**), we simulated the model by varying the meta-parameters c and P_o (**Fig. 3B**). When the agent strongly desired the reward ($P_o = 0.99$, nearly equal to one) with no curiosity ($c = 0$), the agent preferred the right option with a higher reward probability (**Fig. 3C-(a)**). If the agent had no desire for a reward ($P_o = 0.5$) with high curiosity ($c = 10$), the agent preferred the right option with a higher reward probability (**Fig. 3C-(b)**). Although this behavior seems to be rational at first glance, the agent did not seek the reward, but rather sought the information, that is, belief update, driven by curiosity, resulting in preference for the uncertain option. When the agent moderately desired the reward ($P_o = 0.5$ to 1) with negative curiosity ($c = -10$), the agent conservatively selected the same option as the previous selection (**Fig. 3C-(c)**). As a result, the agent continuously selected either of the two options depending on the first selection. These results were expected.

However, we obtained a non-trivial result in another situation, where the reward probabilities were 0.5 for the left and 1 for the right (**Fig. 3A'–C'**). As in the previous situation, the agents with a strong desire for the reward ($P_o = 0.99$, nearly equal to 1) preferred the right option with a higher reward probability (**Fig. 3C'-(a)**). The agent with no desire for reward ($P_o = 0.5$) and high curiosity ($c = 10$) preferred the left option with a lower reward probability (**Fig. 3C'-(b)**). This seemingly irrational behavior was the outcome of focusing on satisfying curiosity and not seeking rewards. In addition, as seen in the previous situation (**Fig. 3C-(c)**), the agents with negative curiosity ($c = -10$), irrespective of the desire for the reward, exhibited conservative selection (**Fig. 3C'-(c)**). Taken together, these results clearly indicate that behavioral patterns largely depend on the degree of conflict between reward and curiosity, which has been difficult to define because of ambiguity (**Fig. 3**).

inverse FEP: Bayesian estimation of internal state

In the above cases, we assumed a constant balance between reward and curiosity. However, in reality, our feelings swing in a context-dependent manner. Although it is important to decipher the temporal swinging of the conflict between reward and curiosity in terms of neuroscience and psychology, it is difficult to quantify the conflict because of its temporal dynamics. Here, we developed a machine learning method called inverse FEP (iFEP) to quantitatively decipher the temporal dynamics of the internal state including curiosity meta-parameter from behavioral data.

In developing iFEP, we needed to switch the viewpoint from the agent to the observer; that is, from animals to us. In the agent-SSM, we described the sequential recognition of reward probabilities by the agent (**Fig. 1C, Fig. 4A,B**). Conversely, to determine the internal state of the agent, e.g., intensity of curiosity c_t, recognition $\lambda_{i,t}$ and its confidence $\gamma_{i,t}$, we developed a state-space model from the observer's eye (observer-SSM) (**Fig. 4B**). In the observer-SSM, the intensity of curiosity was assumed to change continuously over time, and the agent's recognition of the reward probability was updated by equations (1) and (2) following FEP, but they were unknown to the observers. In addition, the agent's actions were assumed to be generated depending on the intensity of its curiosity, recognition, and confidence, as described in equation (3), but the observers can only monitor the agent's action and the presence of a reward. In iFEP, based on the observer-SSM, we estimate the latent internal state of agent z from observation x in a Bayesian manner as follows:

$$P(z_{1:T}|x_{1:T}) \propto P(x_{1:T}|z_{1:T})P(z_{1:T}), \tag{4}$$

where $z_{i,1:T} = \{\mu_{i,1:T}, p_{i,1:T}, c_{1:T}\}, x_{1:T} = \{a_{1:T}, \gamma_{1:T}\}$, and the subscript $1{:}T$ indicates steps 1 to T. In this Bayesian estimation, a posterior distribution $P(z_{1:T}|x_{1:T})$ represents the observer's recognition of the estimated $z_{1:T}$ given observation $x_{1:T}$ with uncertainty. A prior distribution $P(z_{1:T})$ represents our belief, which is expressed as the ReCU model with random motion of the curiosity meta-parameter c as

$$c_t = c_{t-1} + \epsilon\zeta_t,$$

where ζ_t indicates white noise with zero mean and unit variance, and ϵ indicates its noise intensity. The likelihood $P(x_{1:T}|z_{1:T})$ represents the probability that $x_{1:T}$ was observed assuming $z_{1:T}$, which also follows the ReCU model. This Bayesian estimation, namely iFEP, was conducted using a particle filter and Kalman backward algorithm (see Methods).

Validation of iFEP

We tested the validity of the iFEP method by applying it to the artificial data generated by the ReCU model. Specifically, we simulated a model agent with non-constant curiosity in the two-choice task, where reward probabilities varied temporally. We then demonstrated that iFEP estimated the ground truth of the internal state of the simulated agent, that is, the agent's intensity of curiosity, recognition, and confidence (**Fig. 5**). Therefore, iFEP must be powerful in clarifying decision-making processing and the accompanying temporal swing in the conflict between reward and curiosity.

Negative curiosity of rat behaviors decoded by iFEP

Finally, we applied iFEP to actual rat behavioral data from the two-choice task experiment with temporally varying reward probabilities (**Fig. 6A**)[17]. In this experiment, once the reward probabilities were suddenly changed in a discrete manner, the rat slowly adapted to select the option with the higher reward probability (**Fig. 6B**), suggesting that the rat sequentially updated its recognition of the reward probability. Based on these behavioral data of the rat, iFEP estimated the internal state, that is, the intensity of curiosity, the recognized reward probabilities, and their confidence levels (**Fig. 6C-E**). We found that the rat was not perfectly aware of the true reward probabilities but was able to recognize increases and decreases in reward probability (**Fig. 6C and D).** We also found that confidence increased with choice and decreased with no choice (**Fig. 6E and F**). Interestingly, the curiosity held by the rat was always estimated to be negative (**Fig. 6G**). In other words, the rat conservatively preferred certain choices but did not explore uncertain choices, regardless of the recognition of reward probability.

To further show that negative curiosity was not estimated by chance, we conducted Monte Carlo testing (**Fig. S1**). As a null hypothesis, we considered an agent who has no curiosity and decides on a choice only depending on its recognition of the reward probability. Following the null hypothesis, we simulated a series of the agent's choices using iFEP-estimated reward probability recognition and its confidence with zero curiosity (i.e., $c_t = 0$). The same simulations were repeated 1,000 times, and the curiosity was estimated using iFEP for each (**Fig. S1B**). We then plotted the null distribution for temporal average of estimated curiosity, as a test statistic (**Fig. S1B**). Compared to the null distribution, the temporal average curiosity estimated from the actual rat behavior was located to the left of the significance level ($p = 0$). These results suggest that rats have negative curiosity; they made choices based on confidence in reward probability recognition rather than the recognition itself.

Discussion

Previous studies have discussed decision-making subject to rationality and optimality, in which irrational behaviors typical of animals cannot be tractable[18]. By contrast, we proposed the ReCU model, which expresses the mental conflict between reward and curiosity. In addition, we developed the iFEP method to decode the conflict between reward and curiosity based on the time-series behavioral data of a two-choice task.

What is the neural basis that controls psychological states, including mental conflict? This question cannot be addressed simply by comparing neural activities and animal actions because the psychological state is not reflected from each primitive action itself but must be behind a series of actions. Thus, it is important to define psychological states and estimate latent psychological states, such as curiosity, confidence, and reward prediction error, from animal behaviors, which enables us to compare the estimated psychological states and neural

activities[12,15,19]. The iFEP method is an essential tool for this purpose. Therefore, the iFEP approach will contribute to the future development of neuroscience for psychological states.

Our approach has three novel characteristics compared with previous models. First, we addressed the irrationality of decision-making. The expected free energy was originally expressed as the sum of the expected information gain and the expected reward, both of which were equally weighted[10,11], leading to Bayesian optimality. However, animals, including humans, perform decision-making under bounded rationality. Consequently, their computational performance is restricted by the limited size of the brain and time constraints for needing to make quick decisions. Thus, they do not necessarily achieve Bayesian optimization. By contrast, our study formulated irrational decision-making by introducing arbitrary weights on the expected information as a meta-parameter controlling curiosity. Therefore, our model represents individuality. Second, we addressed temporally changing curiosity, which might be situation-dependent. Animals and humans face a dilemma between reward and curiosity, which causes curiosity to vary over time. Third, we addressed an inverse problem using the iFEP method, which quantitatively decodes psychological states, including the intensity of curiosity, from the viewpoint of an observer of animals. This method is able to evaluate not only individual-specific properties, such as curious and conservative characters, but also temporal mental swings in a situation-dependent manner.

Decision-making on the two-choice task has been computationally addressed based on two types of modeling: RL and FEP. Studies based on RL described the temporal update of action values, which is called Q-learning[6], but with a lack of recognition of the environment. In RL reinforcement modeling, it is common to represent the degree of exploration by a temperature parameter, which derives randomness in action selection but does not lead to information-seeking behavior, which corresponds to curiosity in our approach. Several studies have estimated the dynamics and parameters of Q-learning[15,17,19,20]. By contrast, Schwartenbeck et al. modeled decision-making behavior in a two-choice task based on FEP[21]. In their study, they succeeded in representing animal behaviors driven not only by reward but also by curiosity. However, they assumed that the intensity of curiosity was fixed at one ($c_t = 1$), at which the Bayes optimality was satisfied, and did not address its temporal variability as assumed and estimated in our iFEP approach. Furthermore, the related modeling differs from RL and FEP. Ortega and Braun formulated FEP, which describes irrational decision-making[22,23]. Interestingly, their formulation was based on microscopic thermodynamics, and the temperature parameters controlled this irrationality. However, the thermodynamics-based FEP did not treat the sequential update of the recognition from the observations.

Finally, it is worth discussing future perspectives of our iFEP approach in medicine. Generally, mental diagnosis relies on medical interviews and has not been quantitatively evaluated. Our iFEP method can quantitatively estimate the psychological state of patients based on their behavioral data. For example, patients with social withdrawal, also known as "Hikikomori," have no interest in anything. In this case, social withdrawal would be characterized

by a negative value of curiosity in our FEP model. Therefore, the iFEP method could be effective in diagnosing mental disorders.

Methods

FEP for reward probability recognition

In the reward probability recognition process, the agent assumes an environment expressed by a state-space model (**Fig. 1C**) as

$$P(\mathbf{w}_t|\mathbf{w}_{t-1}) = \mathcal{N}(\mathbf{w}_t|\mathbf{w}_{t-1}, \sigma^2\mathbf{I}),$$
$$P(o_t|\mathbf{w}_t, \mathbf{a}_t) = \prod_i \left[f(w_{i,t})^{o_t}\{1 - f(w_{i,t})\}^{1-o_t}\right]^{a_{i,t}}, \quad (5)$$

where $\mathbf{w}_t$ and o_t denote the latent variables controlling the reward probability of both options at step t ($\mathbf{w}_t = (w_{1,t}, w_{2,t})^T$) and the observation of the presence of the reward ($o_t \in \{0,1\}$), respectively; $\mathbf{a}_t$ denotes the agent's action at step t, which is represented by a one-hot vector ($\mathbf{a}_t \in \{(1,0)^T, (0,1)^T\}$); $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes the Gaussian distribution mean $\boldsymbol{\mu}$ and variance Σ; σ^2 denotes the variance of the transition probability of $\mathbf{w}$; and $f(w_{i,t}) = 1/(1 + e^{-w_{i,t}})$, which represents the probability of reward of option i at step t. The initial distribution of $\mathbf{w}_1$ is given by $P(\mathbf{w}_1) = N(\mathbf{w}_1|0, \kappa\mathbf{I})$, where κ denotes variance.

Here, we modeled the agent's recognition process of reward probability using sequential Bayesian updating as follows:

$$P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t}) \propto P(o_t|\mathbf{w}_t, \mathbf{a}_t) \int P(\mathbf{w}_t|\mathbf{w}_{t-1}) P(\mathbf{w}_{t-1}|\boldsymbol{o}_{1:t-1}, \mathbf{a}_{1:t-1}) d\mathbf{w}_{t-1}.$$

Because of the non-Gaussian $P(o_t|\mathbf{w}_t, \mathbf{a}_t)$, the posterior distribution of $\mathbf{w}_t$, $P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t})$, becomes non-Gaussian and cannot be calculated analytically. To avoid this problem, we introduced a simple posterior distribution approximated by a Gaussian distribution, as follows:

$$Q(\mathbf{w}_t|\varphi_t) = \mathcal{N}(\mathbf{w}_t|\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t^{-1}) \fallingdotseq P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t}),$$

where $\varphi_t = \{\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t\}$, and $\boldsymbol{\mu}_t$ and $\boldsymbol{\Lambda}_t$ denote the mean and precision, respectively ($\boldsymbol{\mu}_t = (\mu_{1,t}, \mu_{2,t})^T$; $\boldsymbol{\Lambda}_t = diag(p_{1,t}, p_{2,t})$). $Q(\mathbf{w}_t|\varphi_t)$ denotes the recognition distribution. The model agent aims to update the recognition distribution through φ_t at each time step by minimizing the surprise, which is defined by $-lnP(o_t|o_{1:t-1})$. The surprise can be decomposed as follows:

$$-lnP(o_t|o_{1:t-1}) = \int \frac{Q(\mathbf{w}_t|\varphi_t) lnQ(\mathbf{w}_t|\varphi_t)}{P(o_t, \mathbf{w}_t|o_{1:t-1}, \mathbf{a}_{1:t})} d\mathbf{w}_t - \mathrm{KL}[Q(\mathbf{w}_t|\varphi_t)||P(\mathbf{w}_t|o_{1:t}, \mathbf{a}_{1:t})],$$

where $\mathrm{KL}[q(\mathbf{x})||p(\mathbf{x})]$ denotes the Kullback–Leibler (KL) divergence between the probability distributions $q(\mathbf{x})$ and $p(\mathbf{x})$. Because of the non-negativity of KL divergence, the first term is the upper bound of the surprise:

$$F(o_t, \varphi_t) = \int Q(\mathbf{w}_t|\varphi_t) lnQ(\mathbf{w}_t|\varphi_t) d\mathbf{w}_t + \int Q(\mathbf{w}_t|\varphi_t) E(o_t, \mathbf{w}_t) d\mathbf{w}_t,$$

which is called the free energy, where $E(o_t, \mathbf{w}_t) = -lnP(o_t, \mathbf{w}_t | o_{1:t}, \mathbf{a}_{1:t})$. The first term of the free energy corresponds to the negative entropy of a Gaussian distribution:

$$F_1 = \int Q(\mathbf{w}_t|\varphi_t) lnQ(\mathbf{w}_t|\varphi_t) d\mathbf{w}_t \,.$$

The second term is approximated as

$$\begin{aligned} F_2 &= \int Q(\mathbf{w}_t|\varphi_t) E(o_t, \mathbf{w}_t) d\mathbf{w}_t \\ &\cong \int Q(\mathbf{w}_t|\varphi_t) \left\{ E(o_t, \boldsymbol{\mu}_t) + \frac{dE}{dw}(\mathbf{w}_t - \boldsymbol{\mu}_t) + \frac{1}{2}\frac{d^2E}{dw^2}(\mathbf{w}_t - \boldsymbol{\mu}_t)^2 \right\} d\mathbf{w}_t \\ &= E(o_t, \mathbf{w}_t)|_{w_t=\mu_t} + \frac{1}{2}\frac{d^2E}{dw^2}\bigg|_{w_t=\mu_t} \boldsymbol{\Lambda}_t^{-1} \end{aligned} \tag{6}$$

Note that $E(o_t, \mathbf{w}_t)$ is expanded by a second-order Taylor series around $\boldsymbol{\mu}_t$. At each time step, the agent updates φ_t by minimizing $F(o_t, \varphi_t)$.

Calculation of free energy

The free energy is derived as follows. F_1 simply becomes

$$F_1 = \frac{1}{2} ln2\pi p_{1,t}^{-1} + \frac{1}{2} ln2\pi p_{2,t}^{-1} + const.$$

For computing F_2,

$$\begin{aligned} P(o_t, \mathbf{w}_t | o_{1:t-1}, \mathbf{a}_{1:t}) &= P(o_t|\mathbf{w}_t, \mathbf{a}_t) \int P(\mathbf{w}_t|\mathbf{w}_{t-1}) P(\mathbf{w}_{t-1}|o_{1:t-1}, \mathbf{a}_{1:t-1}) d\mathbf{w}_{t-1} \\ &\cong P(o_t|\mathbf{w}_t, \mathbf{a}_t) \int P(\mathbf{w}_t|\mathbf{w}_{t-1}) \, \mathrm{N}(\mathbf{w}_{t-1}|\boldsymbol{\mu}_{t-1}, \boldsymbol{\Lambda}_{t-1}^{-1}) d\mathbf{w}_{t-1}. \end{aligned}$$

In the second line of this equation, we use the approximated recognition distribution as the previous posterior $P(\mathbf{w}_{t-1}|o_{1:t-1}, \mathbf{a}_{1:t-1})$. This equation can be written as follows:

$$P(o_t, \mathbf{w}_t | o_{1:t-1}, \mathbf{a}_{1:t}) \cong \prod_i \left[f(w_{i,t})^{o_t} \{1 - f(w_{i,t})\}^{1-o_t} \right]^{a_{i,t}} N(w_{1,t}|\mu_{1,t-1}, p_{1,t}^{-1} + \sigma^2) N(w_{2,t}|\mu_{2,t-1}, p_{2,t}^{-1} + \sigma^2).$$

Then,

$$E(o_t, \mathbf{w}_t)|_{w_t=\mu_t} = E_1(o_t, \mu_{1,t}) + E_2(o_t, \mu_{2,t}) + const,$$

where

$$\begin{aligned} E_i(o_t, \mu_{i,t}) = a_{i,t}\left[o_t \log f(\mu_{i,t}) + (1 - o_t) log\{1 - f(\mu_{i,t})\} \right] - \frac{1}{2}\frac{(\mu_{i,t} - \mu_{i,t-1})^2}{p_{i,t}^{-1} + \sigma^2} \\ - \frac{1}{2}\log(p_{i,t}^{-1} + \sigma^2). \end{aligned} \tag{7}$$

Thus, F_2 is calculated by substituting equation (7) into equation (6). Taken together,

$$F(o_t, \varphi_t) = \sum_i \left\{ E_i(o_t, \mu_{i,t}) + \frac{1}{2} \frac{d^2 E_i}{dw_{i,t}^2}\bigg|_{w_{i,t}=\mu_{i,t}} p_{i,t}^{-1} + \frac{1}{2} ln 2\pi p_{i,t}^{-1} \right\}. \quad (8)$$

Sequential updating of the agent's recognition

The updating rule for φ_t was derived by minimizing the free energy. The optimized $p_{i,t}$ can be computed by $\partial F / \partial p_{i,t}^{-1} = 0$, which leads to the following:

$$p_{i,t} = \frac{d^2 E_i}{dw_{i,t}^2}\bigg|_{w_{i,t}=\mu_{i,t}}.$$

By substituting $p_{i,t}$ into equation (8), the second term in the summation becomes constant, irrespective of $\mu_{i,t}$. Thus, $\mu_{i,t}$ is updated by minimizing only the first term as

$$\mu_{i,t} = \mu_{i,t-1} - \alpha \delta_{i,a} \frac{\partial E_i}{\partial \mu_{i,t}}\bigg|_{\mu_{i,t}=\mu_{i,t-1}},$$

where α is the learning rate. These two equations finally lead to equations (1) and (2).

Model for action selection

The agent selects a choice with higher expected utility defined by

$$U(\mathbf{a}_{t+1}) = c_t \cdot E_{P(o_{t+1}|\mathbf{a}_{t+1})}\left[\mathrm{KL}[Q(\mathbf{w}_{t+1}|o_{t+1}, \mathbf{a}_{t+1})||Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})]\right] + E_{P(o_{t+1}|\mathbf{a}_{t+1})}[\ln P(o_{t+1})]. \quad (9)$$

The first and second terms represent the expected information gain and expected reward, respectively, and c_t denotes the intensity of curiosity at time t.

In the first term, the posterior and prior distributions of $\mathbf{w}_{t+1}$ in the KL divergence are derived as

$$Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1}) = \int P(\mathbf{w}_{t+1}|\mathbf{w}_t) Q(\mathbf{w}_t) d\mathbf{w}_t,$$

$$Q(\mathbf{w}_{t+1}|o_{t+1}, \mathbf{a}_{t+1}) = \frac{P(o_{t+1}|\mathbf{w}_{t+1}, \mathbf{a}_{t+1}) Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})}{P(o_{t+1}|\mathbf{a}_{t+1})},$$

respectively. This KL divergence is called a Bayesian surprise, representing the extent to which the agents' beliefs are updated by observation. Note that o_{t+1} is a future observation, therefore the first term was expected by $P(o_{t+1}|\mathbf{a}_{t+1})$, which can be calculated by

$$P(o_{t+1}|\mathbf{a}_{t+1}) = \int P(o_{t+1}|\mathbf{w}_{t+1}, \mathbf{a}_{t+1}) Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1}) d\mathbf{w}_{t+1}.$$

In the second term, the reward is quantitatively interpreted as the desired probability of o_{t+1}. For the two-choice task, we use

$$P(o_{t+1}) = P_o^{o_{t+1}} (1 - P_o)^{(1-o_{t+1})},$$

where Po indicates the desired probability of the presence of a reward. According to the probabilistic interpretation of reward[10,11], the presence and absence of a reward can be evaluated

by $\ln P_o$ and $\ln(1-P_o)$, respectively. Because rewards are relative, in this study, we set $\ln\{P_o/(1-P_o)\}$ and 0 for the presence and absence of a reward, respectively.

Calculation of the expected utility

Here, we present the calculation of the expected utility. The KL divergence in the first term of equation (9) can be transformed into

$$E_{p(o_{t+1}|\mathbf{a}_{t+1})}\left[KL[Q(\mathbf{w}_{t+1}|o_{t+1},\mathbf{a}_{t+1})||Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})]\right] = \mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) - \mathrm{H}(o_{t+1}),$$

where the first and second terms represent the conditional and marginal entropies, respectively:

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) = \mathrm{E}_{P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})}[-\ln P(o_{t+1}|\mathbf{w}_{t+1},\mathbf{a}_{t+1})], \tag{10}$$

$$\mathrm{H}(o_{t+1}) = \mathrm{E}_{P(o_{t+1}|\mathbf{a}_{t+1})}[-\ln P(o_{t+1}|\mathbf{a}_{t+1})].$$

The conditional entropy $\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1})$ can be calculated by substituting equation (5) into equation (10) as

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) = -E_{Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})}\left[\sum_i a_{i,t+1} g(w_{i,t+1})\right],$$

where

$$g(w) = f(w)\ln f(w) + (1-f(w))\ln(1-f(w)).$$

Here, we approximately calculate this equation by using the second-order Taylor expansion as

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) \cong -E_{Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1})}\left[\sum_i a_{i,t+1}\left\{g(\mu_{i,t+1}) + \frac{\partial g}{\partial w_{i,t+1}}(w_{i,t+1}-\mu_{i,t+1}) + \frac{1}{2}\frac{\partial^2 g}{\partial w_{i,t+1}^2}\left(w_{i,t+1}-\mu_{i,t+1}\right)^2\right\}\right],$$

which leads to

$$\mathrm{H}(o_{t+1}|\mathbf{w}_{t+1}) = \sum_i a_{i,t+1}\left[f(\mu_{i,t+1})\ln f(\mu_{i,t+1}) + \left(1-f(\mu_{i,t+1})\right)\ln\left(1-f(\mu_{i,t+1})\right) + \frac{1}{2}\left\{f(\mu_{i,t+1})\left(1-f(\mu_{i,t+1})\right)\left(1+\left(1-2f(\mu_{i,t+1})\right)\ln\frac{f(\mu_{i,t+1})}{1-f(\mu_{i,t+1})}\right)\right\}(p_{i,t}^{-1}+\alpha)\right].$$

The marginal entropy $\mathrm{H}(o_{t+1})$ can be calculated as

$$\mathrm{H}(o_{t+1}) = \sum_i a_{i,t+1}\{P(o_{t+1}=0|\mathbf{a}_{t+1})\ln P(o_{t+1}=0|\mathbf{a}_{t+1}) + P(o_{t+1}=1|\mathbf{a}_{t+1})\ln P(o_{t+1}=1|\mathbf{a}_{t+1})\},$$

where

$$P(o_{t+1}|\mathbf{a}_{t+1}) = \int P(o_{t+1}|\mathbf{w}_{t+1}, \mathbf{a}_{t+1}) Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1}) d\mathbf{w}_{t+1}$$

$$= \int \prod_i \left\{ f\left(w_{i,t+1}\right)^{o_{t+1}} \left(1 - f\left(w_{i,t+1}\right)\right)^{1-o_{t+1}} \right\}^{a_{i,t+1}} Q(\mathbf{w}_{t+1}|\mathbf{a}_{t+1}) d\mathbf{w}_{t+1}$$

$$= \prod_i \Big[f(\mu_{i,t+1})^{o_{t+1}} \left(1 - f(\mu_{i,t+1})\right)^{1-o_{t+1}}$$

$$+ 1^{o_{t+1}} (-1)^{1-o_{t+1}} \frac{1}{2} f(\mu_{i,t+1}) \left\{ 1 - f(\mu_{i,t+1}) \left\{ 1 - 2f(\mu_{i,t+1}) \right\} \right\} (p_{i,t}^{-1} + \alpha) \Big]^{a_{i,t+1}}.$$

The second term of the expected utility (equation (10)) is calculated as

$$E_{P(o_{t+1}|\mathbf{a}_{t+1})}[\ln P(o_{t+1})] = E_{P(o_{t+1}|\mathbf{a}_{t+1})}[o_{t+1} \ln P_o + (1 - o_{t+1}) \ln (1 - P_o)]$$
$$= P(o_{t+1} = 0|\mathbf{a}_{t+1}) \ln (1 - P_o) + P(o_{t+1} = 1|\mathbf{a}_{t+1}) \ln (1 - P_o).$$

The observer-SSM

We constructed the observer-SSM, which describes the temporal transitions of the latent internal state of agent *z* and the generation of action, from the viewpoint of the observer of the agent. This is depicted as a graphical representation in Fig. 4. As prior information, we assumed that the agent acts based on the internal state, that is, the intensity of curiosity, the recognized reward probabilities, and their confidence levels. The intensity of curiosity was assumed to change temporally as a random walk:

$$c_t = c_{t-1} + \epsilon \zeta_t,$$

where ζ_t denotes white noise with zero mean and unit variance, and ϵ denotes its noise intensity. Other internal states, that is, μ_i and p_i, were assumed to update as equations (1) and (2). The transition of the internal state is expressed by the probability distribution,

$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t|\mathbf{F}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}), \mathbf{\Gamma}), \tag{11}$$

$$\mathbf{F}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \begin{bmatrix} 0 \\ h(\mu_{1,t-1}, p_{1,t-1}, o_t, a_1) \\ h(\mu_{2,t-1}, p_{2,t-1}, o_t, a_2) \\ k(\mu_{1,t-1}, p_{1,t-1}, a_1) \\ k(\mu_{2,t-1}, p_{2,t-1}, a_2) \end{bmatrix},$$

$$\mathbf{\Gamma} = \epsilon^2 \Delta_{1,1},$$

where $\mathbf{z}_t = (c_t, \boldsymbol{\mu}_t^T, \boldsymbol{p}_t^T)^{\mathrm{T}}$; $h(\mu_{i,t-1}, p_{i,t-1}, o_t, a_i)$ and $k(\mu_{i,t-1}, p_{i,t-1}, a_i)$ represent the right-hand sides of equations (1) and (2), respectively, and $\Delta_{i,j}$ denotes a variance-covariance matrix with 1 at (i, j) component and 0 at others. In addition, the agent was assumed to select an action $\mathbf{a}_{t+1}$ based on the expected utilities, as follows:

$$P(\mathbf{a}_{t+1}) = \frac{\exp\left(U(\mathbf{a}_{t+1})\right)}{\sum_{\mathbf{a}} \exp\left(U(\mathbf{a})\right)},$$

and the reward was obtained by the following probability distribution:

$$P(o_t|\mathbf{a}_t) = \prod_i \left\{ \lambda_{i,t}^{o_t} (1 - \lambda_{i,t})^{1-o_t} \right\}^{a_{i,t}}.$$

iFEP by particle filter and Kalman backward algorithm

Based on the observer-SSM, we estimated the posterior distribution of the latent internal state of agent z_t given all observations from 1 to T ($x_{1:T}$) in a Bayesian manner, i.e., $P(z_t|x_{1:T})$. This estimation was done by forward and backward algorithms, which are called filtering and smoothing, respectively.

In filtering, the posterior distribution of z_t given observations until t ($x_{1:t}$) is sequentially updated in a forward direction as

$$P(\mathbf{z}_t|\mathbf{x}_{1:t}) \propto P(\mathbf{x}_t|\mathbf{z}_t,\theta)\int P(\mathbf{z}_t|\mathbf{z}_{t-1},\theta)P(\mathbf{z}_{t-1}|\mathbf{x}_{1:t-1})d\mathbf{z}_{t-1},$$

where $\mathbf{x}_t = (\mathbf{a}_t^T, o_t)^{\mathrm{T}}$ and $\theta = \{\sigma^2, \alpha, P_o\}$. The prior distribution of $\mathbf{z}_1$ is

$$P(\mathbf{z}_1) = \mathcal{N}(\boldsymbol{\mu}_1|\boldsymbol{\mu}_0, \sigma_\mu^2\mathbf{I})\mathcal{N}(\mathbf{p}_1|\mathbf{p}_o, \sigma_p^2\mathbf{I})\mathcal{N}(\mathbf{c}_1|\mathbf{c}_o, \sigma_c^2\mathbf{I}),$$

where $\boldsymbol{\mu}_0, \mathbf{p}_o$and $\mathbf{c}_0$ denote means, and σ_μ^2, σ_p^2 and σ_c^2 denote variances. We used a particle filter[24] to sequentially calculate the posterior $P(\mathbf{z}_t|\mathbf{x}_{1:t})$, which cannot be analytically derived because of the nonlinear transition probability.

After the particle filter, the posterior distribution of $\mathbf{z}_t$ given all observations ($\mathbf{x}_{1:T}$) is sequentially updated in a backward direction as

$$P(\mathbf{z}_t|\mathbf{x}_{1:T}) = \int P(\mathbf{z}_{t+1}|\mathbf{x}_{1:t+1})P(\mathbf{z}_{t+1}|\mathbf{z}_t,\theta)d\mathbf{z}_{t-1}\,. \tag{12}$$

However, this integration is intractable due to non-Gaussian $P(\mathbf{z}_{t-1}|\mathbf{x}_{1:t-1})$, which was represented by particle ensemble in the particle filter, and non-linear relationship between $\mathbf{z}_{t-1}$ and $\mathbf{z}_t$ in $P(\mathbf{z}_t|\mathbf{z}_{t-1},\theta)$ (equation (11)). Thus, we approximated $P(\mathbf{z}_t|\mathbf{x}_{1:t})$ as $\mathcal{N}(\mathbf{z}_t|\mathbf{m}_t, \mathbf{V}_t)$, where $\mathbf{m}_t$ and $\mathbf{V}_t$ denote sample mean and sample variance of particles at t, whereas we linearized $P(\mathbf{z}_t|\mathbf{z}_{t-1},\theta)$ as

$$P(\mathbf{z}_t|\mathbf{z}_{t-1},\theta) \cong \mathcal{N}(\mathbf{z}_t|\mathbf{A}\mathbf{z}_{t-1} + \mathbf{b}, \boldsymbol{\Gamma}),$$

$$\mathbf{A} = \left.\frac{\partial\mathbf{F}(\mathbf{z}_{t-1},\mathbf{a}_{t-1})}{\partial\mathbf{z}_{t-1}}\right|_{\mathbf{m}_t},$$

$$\mathbf{b} = F(\mathbf{m}_t,\mathbf{a}_{t-1}) - \mathbf{A}\mathbf{m}_t,$$

where $\mathbf{A}$ denotes Jacobian matrix. Because these approximations make integration of equation (12) tractable, the posterior distribution $P(\mathbf{z}_t|\mathbf{x}_{1:T})$ can computed by Gaussian distribution as

$$P(\mathbf{z}_t|\mathbf{x}_{1:T}) = \mathcal{N}\left(\mathbf{z}_t|\widehat{\mathbf{m}}_t, \widehat{\mathbf{V}}_t\right).$$

Its mean and variance was analytically updated by Kalman backward algorithms[25]:

$$\widehat{\mathbf{m}}_t = \mathbf{m}_t + \mathbf{J}_t\{\widehat{\mathbf{m}}_{t+1} - (\mathbf{A}\mathbf{m}_t + \mathbf{b})\},$$

$$\widehat{\mathbf{V}}_t = \mathbf{V}_t + \mathbf{J}_t\{\widehat{\mathbf{m}}_{t+1} - (\mathbf{A}\mathbf{V}_t\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Gamma})\}\mathbf{J}_t^{\mathrm{T}},$$

where

$$\mathbf{J}_t = \mathbf{V}_t\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{V}_t\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Gamma})^{-1}.$$

Estimation of parameter in iFEP

To estimate parameter θ, we extended the observer-SSM to a self-organizing state-space model[26] in which θ was also addressed as constant latent variables:

$$P(\mathbf{z}_t, \theta|\mathbf{x}_{1:t}) \propto P(\mathbf{x}_t|\mathbf{z}_t) \int P(\mathbf{z}_t|\mathbf{z}_{t-1}, \theta) P(\mathbf{z}_{t-1}, \theta|\mathbf{x}_{1:t-1}) d\mathbf{z}_{t-1},$$

$$P(\mathbf{z}_t, \theta) = P(\mathbf{z}_1) P(\theta),$$

where $P(\theta) = \mathcal{N}(\sigma^2|m_\sigma, v_\sigma)\mathcal{N}(\alpha|m_\alpha, v_\alpha)\mathcal{N}(P_o|m_{p_o}, v_{p_o})$. To sequentially calculate the posterior $P(\mathbf{z}_t, \theta|\mathbf{x}_{1:t})$ using the particle filter, we augmented the state vector of all particles by adding parameter θ, which was not updated from randomly sampled initial values.

Statistical testing with Monte Carlo simulations

In Fig. S1, we statistically tested negative curiosity estimated in Fig. 6. A null hypothesis is that an agent who has no curiosity (i.e., $c_t = 0$) and decides on a choice only depending on its recognition of the reward probability. Under the null hypothesis, model simulations were repeated 1,000 times under the same experimental conditions as in Fig. 6 and the curiosity was estimated for each using iFEP. We adopted temporal average of estimated curiosity as a test statistic and plotted the null distribution of the test statistic. Compared with the estimated curiosity of the rat behavior, we then computed p-value for one sided left tailed test.

Acknowledgments

We are grateful to Prof. Kenji Doya and Dr. Makoto Ito for providing rat behavioral data. We thank organizers of the tutorial on the free-energy principle in 2019, that inspired this research. This study was supported in part by Grant-in-Aid for Transformative Research Areas (B) [grant number 21H05170], AMED [Grant Number JP21wm0425010], Moonshot R&D–MILLENNIA Program [grant number JPMJMS2024-9] by JST, the Cooperative Study Program of Exploratory Research Center on Life and Living Systems (ExCELLS) [program number 21-102], and the grant of Joint Research by the National Institutes of Natural Sciences [NINS Programme No. 01112102].

Author contributions

H.N. conceived of the project. Y. K. and H.N. developed the method and Y.K. implemented the model simulation. Y.K. and H.N. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

We used the rat behavioral data published by Ito and Doya, *J Neurosci* (2009)[17], which is publicly available at Prof. Doya Kenji's homepage: https://groups.oist.jp/ja/ncu/data. Source codes will be open after publication.

References

1. Helmholtz, H. Handbuch der Physiologischen Optik. 1867 (1867).
2. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
3. Millett, J. D. & Simon, H. A. Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization. *Polit. Sci. Q.* **62**, 621 (1947).
4. Dubey, R. & Griffiths, T. L. Understanding exploration in humans and machines by formalizing the function of curiosity. *Curr. Opin. Behav. Sci.* **35**, 118–124 (2020).
5. Kidd, C. & Hayden, B. Y. The Psychology and Neuroscience of Curiosity. *Neuron* **88**, 449–460 (2015).
6. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).
7. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 815–836 (2005).
8. Friston, K., Kilner, J. & Harrison, L. A free energy principle for the brain. *J. Physiol. Paris* **100**, 70–87 (2006).
9. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
10. Friston, K. *et al.* Active inference and epistemic value. *Cogn. Neurosci.* **6**, 187–214 (2015).
11. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active Inference: A Process Theory. *Neural Comput.* **29**, 1–49 (2017).
12. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science (80-.).* **275**, 1593–1599 (1997).
13. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
14. Tanaka, C. S. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.* **7**, 887–893 (2004).
15. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Neuroscience: Representation of action-specific reward values in the striatum. *Science (80-.).* **310**, 1337–1340 (2005).
16. Ishii, S., Yoshida, W. & Yoshimoto, J. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Networks* **15**, 665–687 (2002).

17. Ito, M. & Doya, K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.* **29**, 9861–9874 (2009).
18. Montague, P. R., Hyman, S. E. & Cohen, J. D. Computational roles for dopamine in behavioural control. *Nature* **431**, 760–767 (2004).
19. Mizoguchi, H. *et al.* Insular neural system controls decision-making in healthy and methamphetamine-treated rats. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3930–E3939 (2015).
20. Katahira, K. The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* **87**, 31–45 (2018).
21. Schwartenbeck, P. *et al.* Computational mechanisms of curiosity and goal-directed exploration. *Elife* **8**, 1–45 (2019).
22. Ortega, P. A. & Braun, D. A. Thermodynamics as a theory of decision-making with information-processing costs Subject Areas : Author for correspondence : *Proc. R. Soc. London. Part A* **469**, 20120683 (2013).
23. Gottwald, S. & Braun, D. A. *The two kinds of free energy and the Bayesian revolution*. *PLoS Computational Biology* vol. 16 (2020).
24. Kitagawa, G. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *J. Comput. Graph. Stat.* **5**, 1–25 (1996).
25. Bishop, C. M. *Pattern recognition and machine learning*. (New York : Springer, [2006] ©2006).
26. Journal, S., Statistical, A. & Sep, N. A Self-Organizing State-Space Model Author (s): Genshiro Kitagawa Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association Stable URL : http://www.jstor.org/stable/2669862 All use subject to http://about.jstor.org/terms. *Am. Stat.* **53**, 326–331 (1999).

Figures

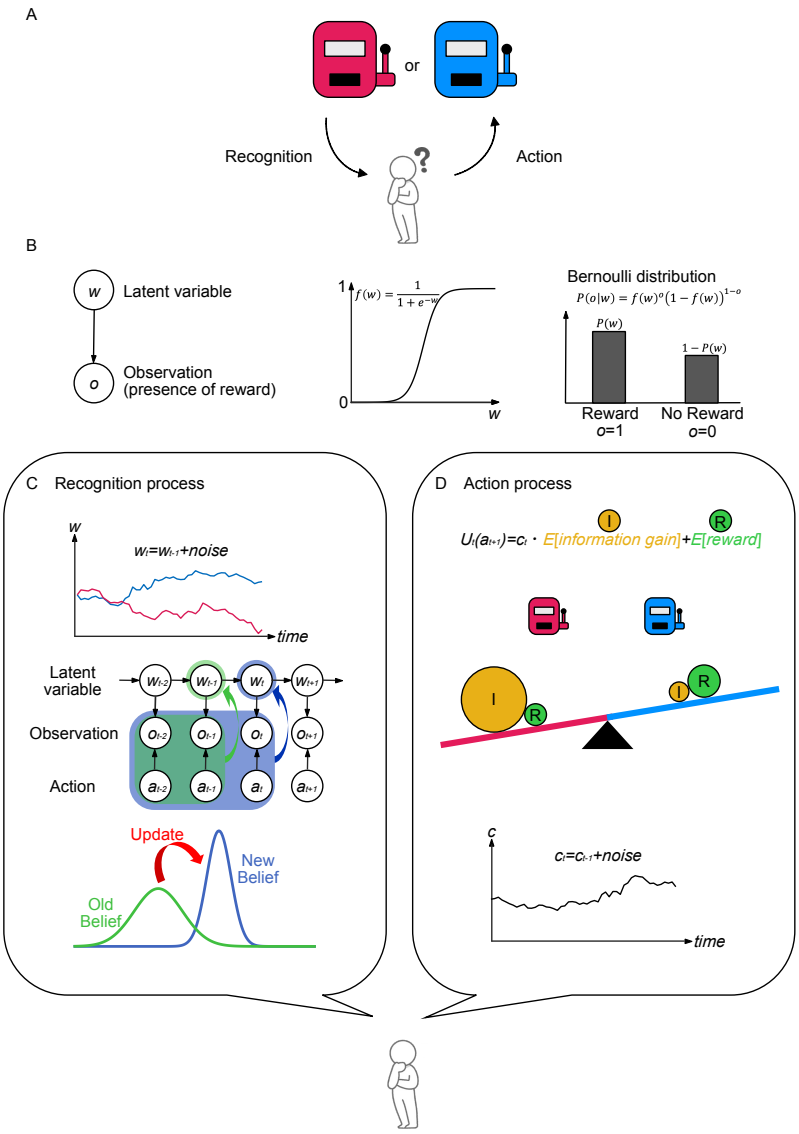

Fig. 1: Decision-making model for the two-choice task with reward–curiosity dilemma
(A) Decision-making in the two-choice task. Reward is provided at different probabilities for each option. The agent does not know those probabilities. Through repeated trial and error, the agent recognizes the world by inferring the latent reward probability of each option, and decides to choose the next action, i.e., option, based on its own inference. **(B)** Generative model of reward. The agent assumes that reward is generated by the latent variable w (left panel), and the reward probability is sigmoidally controlled by w (middle panel). The presence of the reward obeys a Bernoulli distribution (right panel). **(C)** Recognition process by agent. The agent assumes that the reward probabilities change over time owing to fluctuation of the latent variable w_t (upper). The probabilistic dependence among the latent variable w, action a, and the presence of reward o is graphically described (middle). The agent recognizes w as a probability distribution, whose precision, i.e., inverse of variance, represents confidence, and sequentially updates its recognized distribution from previous rewards and actions in a Bayesian manner (lower). **(D)** Action selection process by agent. The agent evaluates the utility of each action using the weighted sum of the expected reward and information gain (upper). The agent compares the utilities for both actions and prefers the option with the larger utility (middle). The expected information gain can be calculated as the difference between the currently recognized distribution and the predicted recognized distribution updated by the future observation (lower).

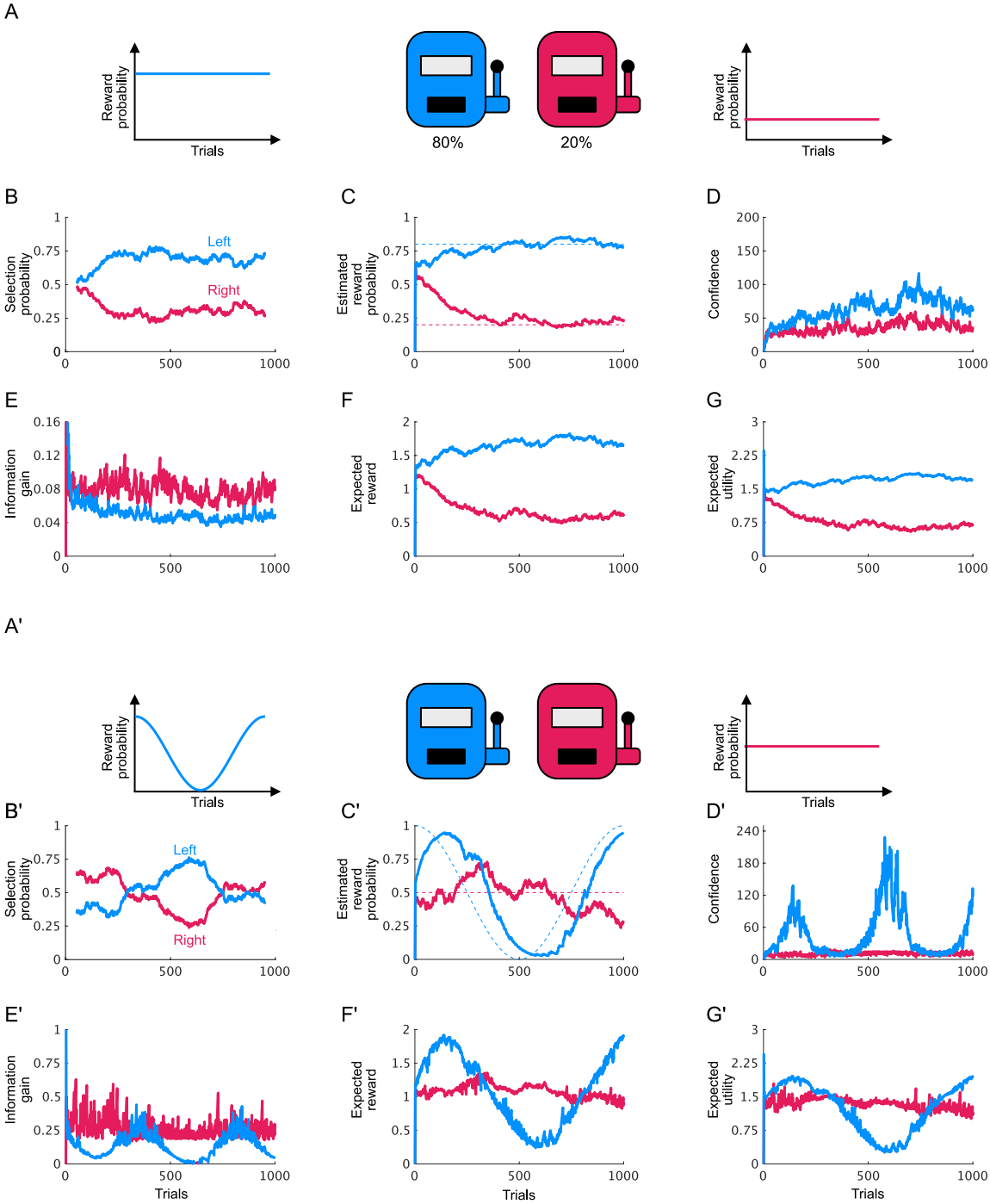

Fig. 2: Simulations of the decision-making model

(A) Two-choice task with constant reward probabilities. **(B)** Selection probabilities for left and right options. These were plotted by moving average 101 window width. **(C)** Recognized reward probabilities for left and right options. Dashed lines represent the ground truths of the reward probabilities. **(D)** Confidences of reward probability recognitions for left and right options. Expected brief updates **(E)**, expected reward **(F)**, and expected utility **(G)** for left and right options. **(A')** Two-choice task with constant and temporally varying reward probabilities for left and right options, respectively. **(B'-G')** The same as (B-G). In these simulations, the curiosity parameter and reward seeking parameter were all kept constant at $c = 1$ and $P_o = 0.9$, respectively.

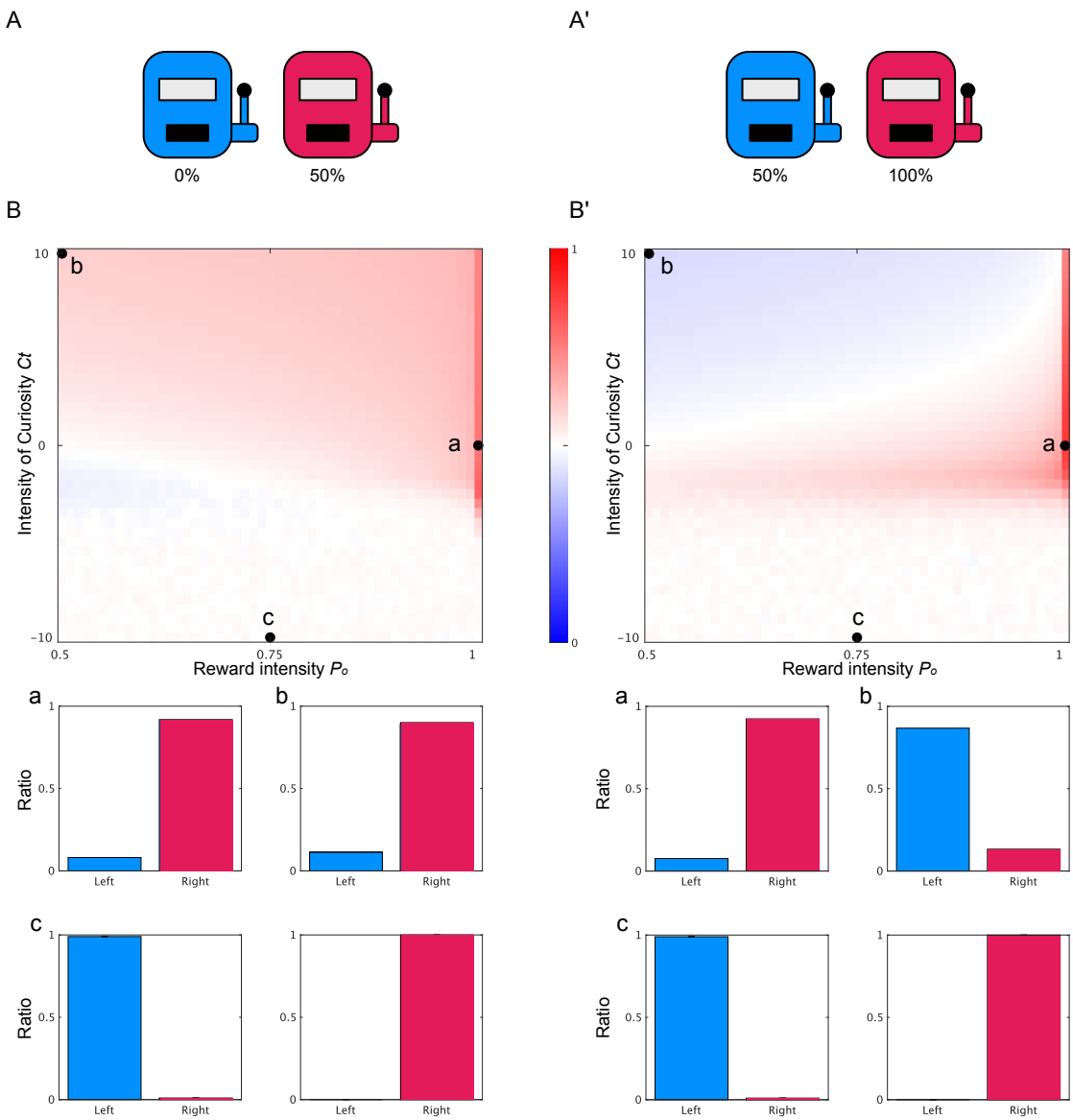

Fig. 3: Curiosity-dependent irrational behaviors

(A) Two-choice task with different, constant reward probabilities: 0% for the left, and 50% for the right. **(B)** Heatmap of selection probability of the right option varying the parameters of curiosity and reward intensity (upper panel). There are three representative conditions indicated by black dots: $c = 0, P_o = 1$ [a], $c = 10, P_o = 0.5$ [b], and $c = -10, P_o = 0.75$ [c]. Lower four panels present the selection ratios of the right option for conditions [a], [b], and [c]. **(A')** Two-choice task with different, constant reward probabilities: 50% for the left and 100% for the right. **(B')** is the same as (B).

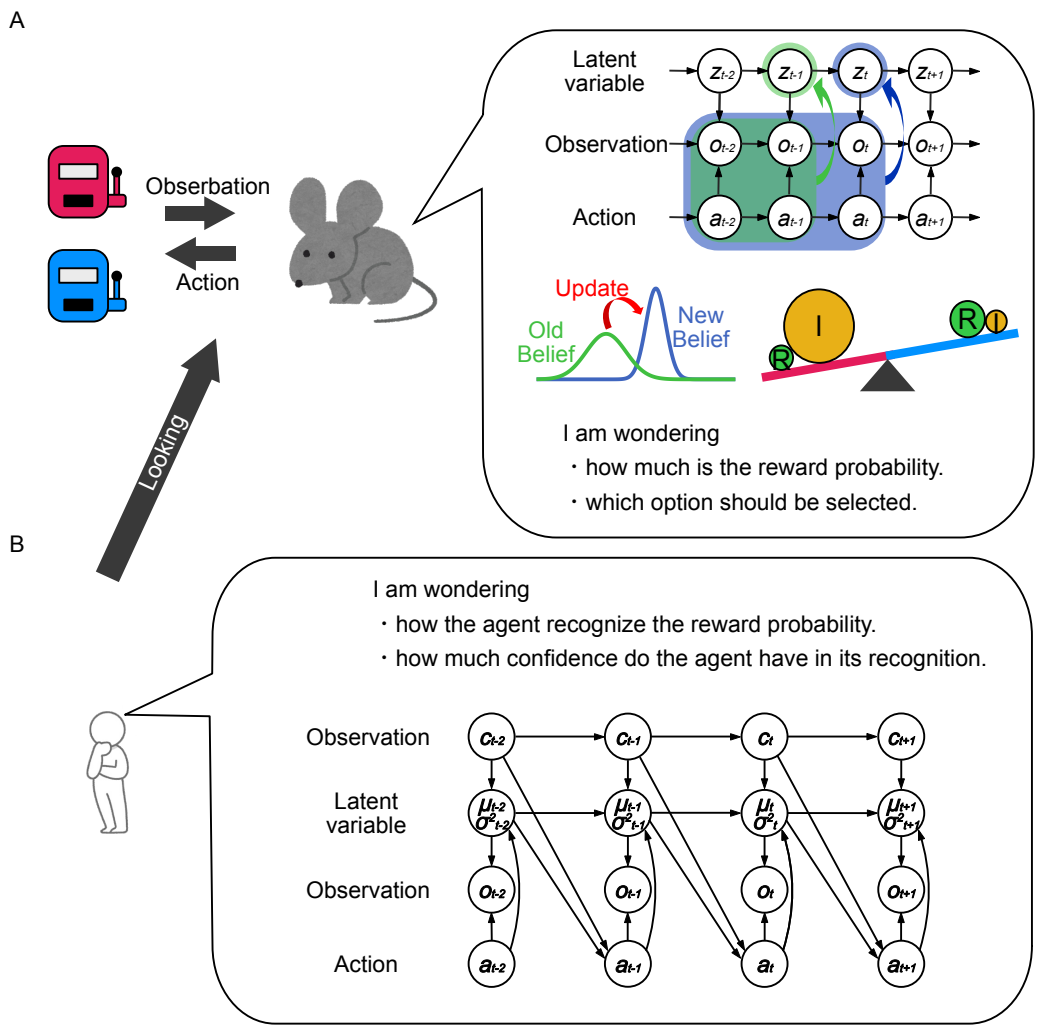

Fig. 4: Scheme of inverse FEP by an observer of a decision-making agent.

(A) An agent performing a two-choice task from the observer's perspective. **(B)** Observer's assumption about the agent's decision-making. The agent is assumed to follow the decision-making model, as described in Fig. 1. **(C)** State-space model of the observer's eye. For the observer, the agent's reward–curiosity conflict, recognized reward probabilities, and their uncertainties are temporally varying latent variables, whereas the agent's action and the presence/absence of a reward are observable. The observer estimates the latent internal states of the agent.

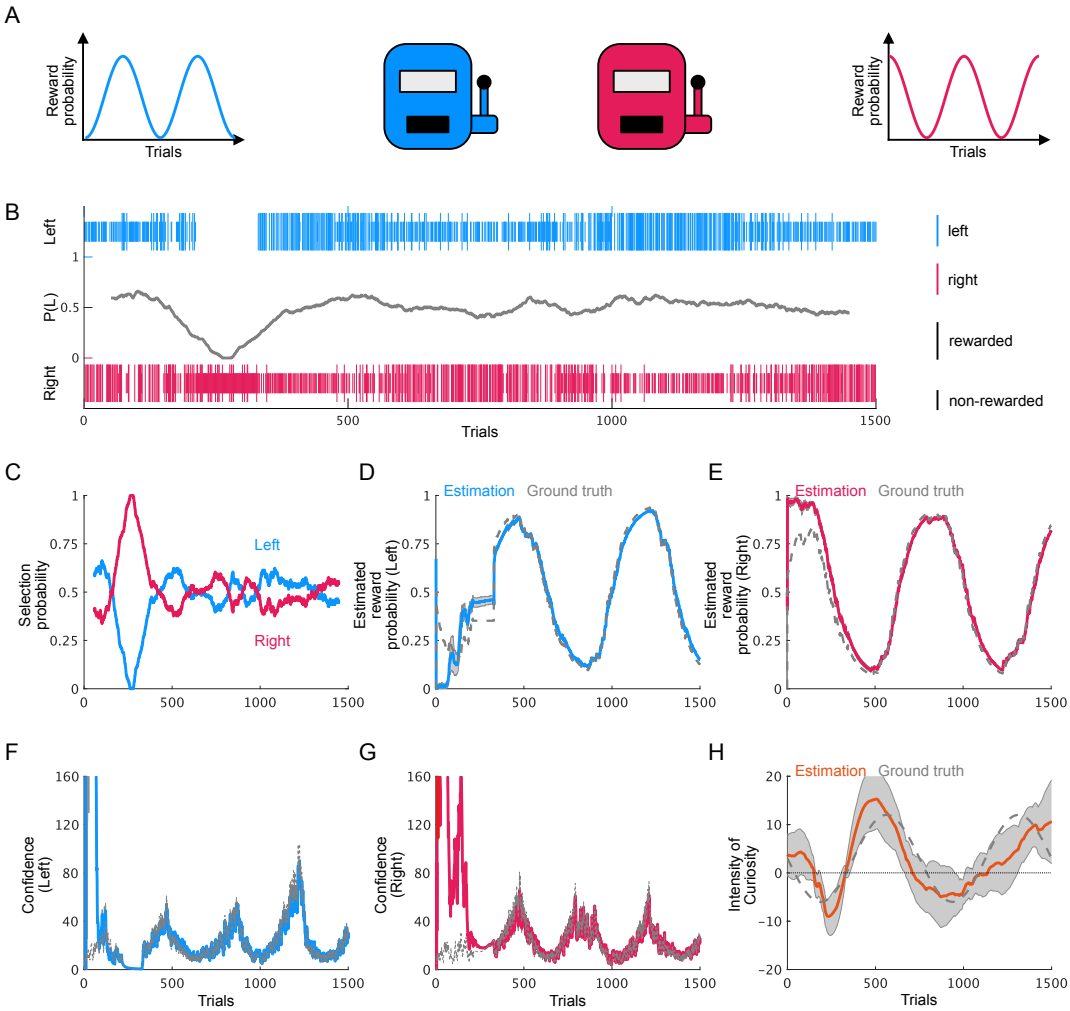

Fig. 5: Estimation of the simulated agent's internal state by inverse FEP

(A) Two-choice task with temporally varying reward probabilities and agent having temporally varying curiosity. **(B)** Simulated agent's behaviors. Blue and red vertical lines indicate selections of left and right options, respectively. Long and short vertical lines indicate the presence and absence of rewards, respectively. Gray line indicates the moving average of selection probability of the left option with 101 window width. **(C)** Moving average of selection probabilities for left and right options. **(D–H)** Simulated agent's behavior-driven estimations of agent-recognized reward probabilities for left (D) and right (E) options, agent's confidence about recognized reward probabilities for left (F) and right (G) options, and agent's curiosity (H). Orange and dashed blue lines represent the estimations and ground truths, respectively.

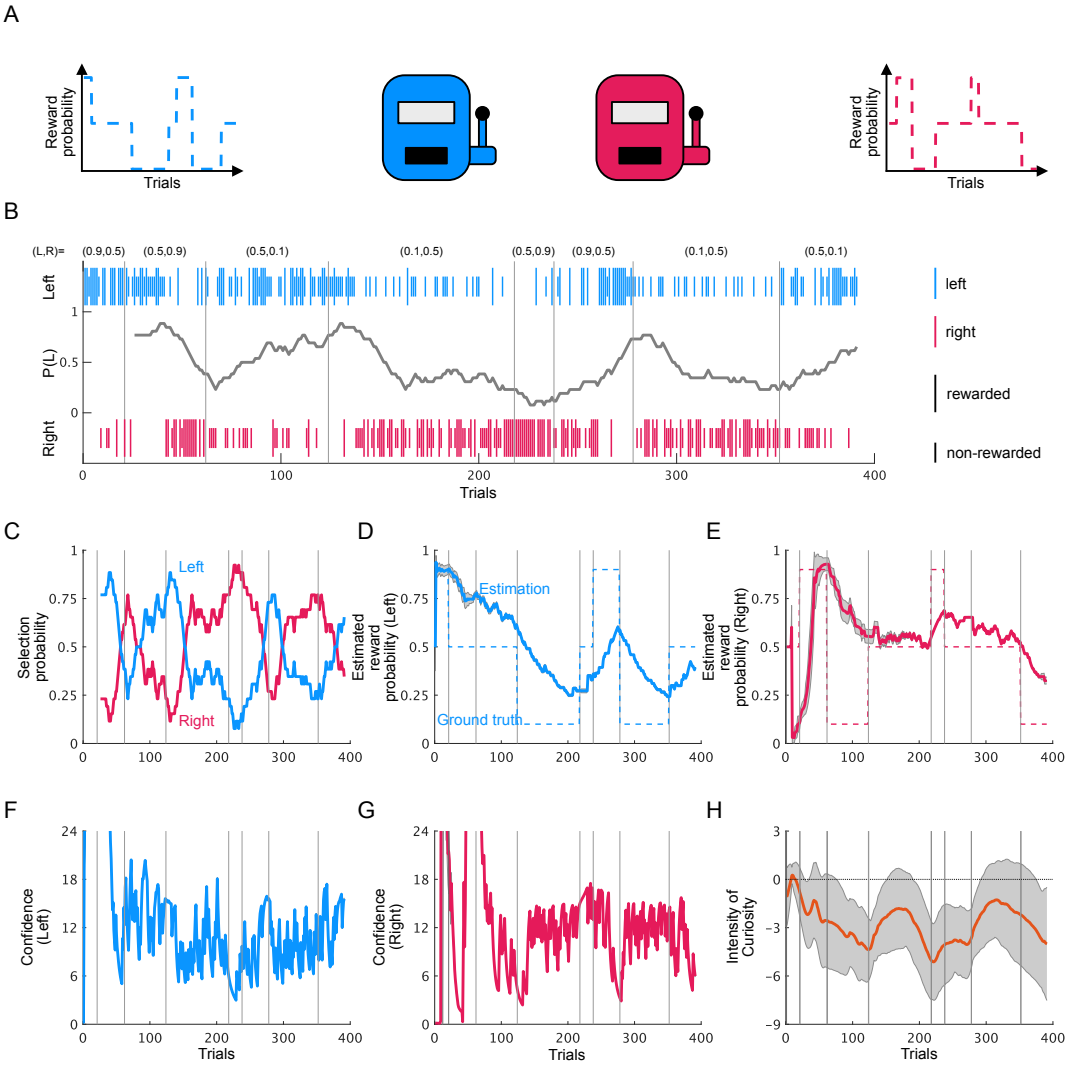

Fig. 6: Estimation of the rat's internal state by inverse FEP

(A) A rat in the two-choice task with temporally switching reward probabilities. **(B)** Rat's behaviors in the two-choice task. A public dataset at https://groups.oist.jp/ja/ncu/data was used[5]. Blue and red vertical lines indicate selections of left and right options, respectively. Long and short vertical lines indicate the presence and absence of rewards, respectively. Gray line indicates the moving average of the selection probability of the left option with 25 window-width backward. **(C)** Moving average of selection probabilities for left and right options. **(D, E)** Rat behavior-driven estimations of agent-recognized reward probabilities for left (D) and right (E) options. Red and dashed blue lines represent the estimations and ground truths, respectively. **(F–H)** Rat behavior-driven estimations of agent's confidence about recognized reward probabilities for left (F) and right (G) options, and agent's curiosity (H).

Supplementary Information

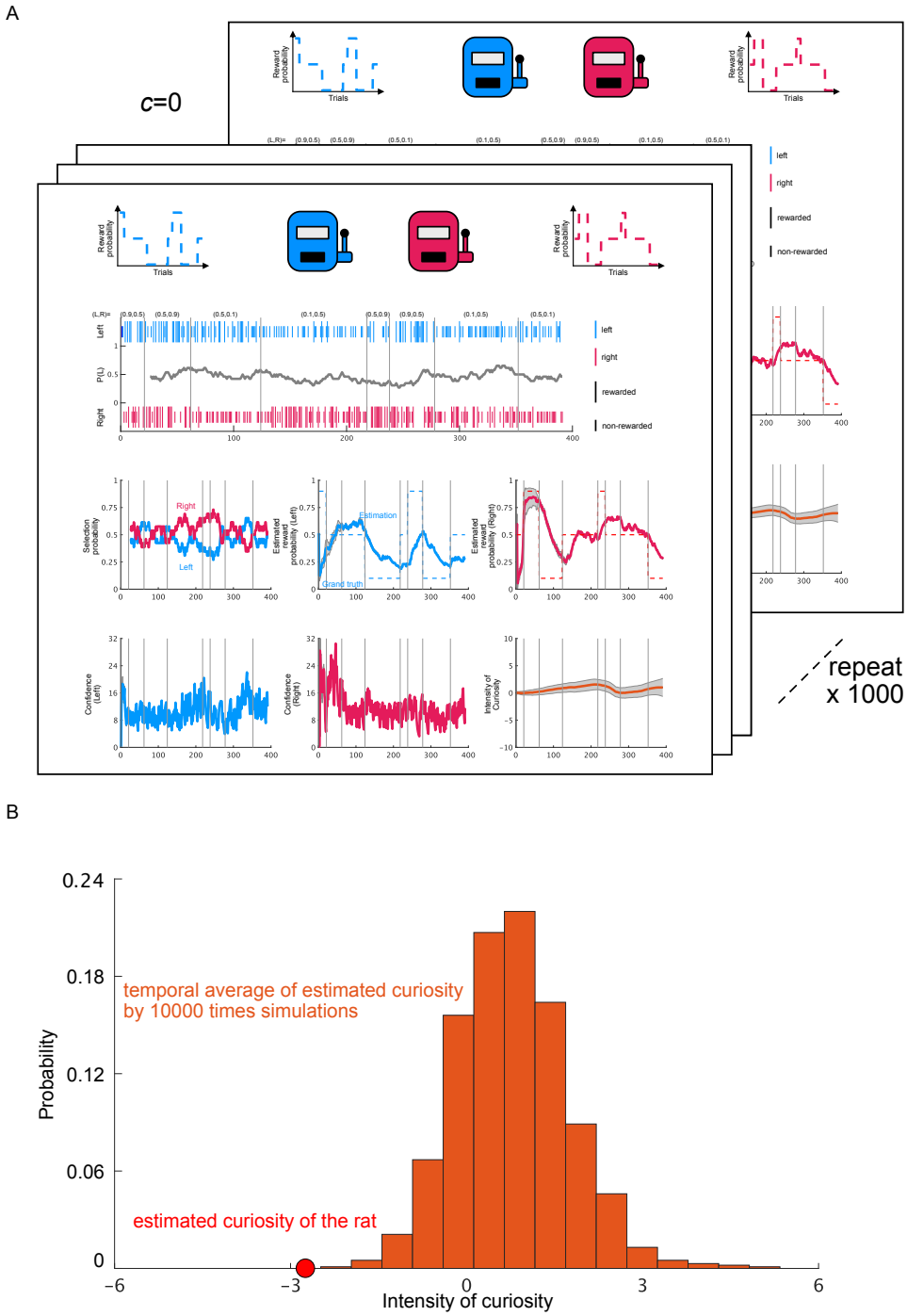

Fig. S1: Monte Carlo statistical testing of estimated negative curiosity in rat

(A) Model simulations conducted under the same experimental conditions as in Fig. 6 under a null hypothesis that the rat behavior is generated by zero curiosity. Simulations were repeated 1,000 times, and the curiosity was estimated using iFEP for each.

(B) Null distribution for temporal average of estimated curiosity by 1,000 times simulations. The estimated curiosity of the rat shown in red circles was on the left side and did not overlap with the distribution, which means p=0 (one-sided, left tailed test).