Estimating Task Representations in a Multidimensional Bandit Task: A Method for Revealing Meta-level Learning

Yuki Tsukamura (tsukkacogsci@gmail.com)[1,2] **& Kazuhiro Ueda (ueda@g.ecc.u-tokyo.ac.jp)**[3]
[1]Graduate School of Education, The University of Tokyo
[2]Japan Society for the Promotion of Science
[3]Graduate School of Arts and Sciences, The University of Tokyo

Abstract

Humans must learn not only which options pay off, but also which features are worth learning about. Yet such meta-level task representations are difficult to observe directly. We propose an inverse inference approach for estimating an individual's hypothesis space over feature dimensions in a multidimensional bandit task. Building on a model previously supported when the hypothesis space is known, we instantiate the model for each candidate hypothesis space and identify the space that best predicts each participant's choice sequence using hierarchical Bayesian estimation and WAIC-based model comparison. In an online experiment, participants' choice accuracy improved both within games and across games, suggesting adaptation at multiple timescales. Moreover, the inferred spaces shifted on average toward the true task structure. These results indicate that changes in task representations are associated with improvements in structured reward learning, and provide a measurement framework for developing process-level models of meta-level learning.

Keywords: reinforcement learning; Bayesian modeling; decision-making; multidimensional bandit task; hypothesis space

Introduction

How can one identify the best restaurant among many options? How can one find the most enjoyable video or song? In such situations, selecting from a large set of alternatives yields subjective experiences of tastiness or enjoyment; based on these outcomes, one may subsequently choose options expected to be better, or instead explore options not previously selected. This cycle of repeated choice, reward-based behavioral updating, and the goal of maximizing (cumulative) reward can be formalized as *bandit tasks*. Because bandit tasks provide a powerful abstraction of learning under repeated decision-making, they have been studied both as algorithmic problems in machine learning (Auer et al., 2002; Lattimore & Szepesvári, 2020) and as behavioral tasks to investigate human learning and choice in psychology (Daw et al., 2006; Steyvers et al., 2009).

In conventional multi-armed bandit formulations, all arms are assumed to be independent, therefore observations from one arm provide no information about the reward distribution of any other arm. In realistic environments, by contrast, subjective restaurant quality is often partially predictable from externally observable attributes, including cuisine category, price bracket, aesthetic cleanliness, patron density, and review-site star ratings. Decision-making in such environments is therefore characterized by generalization through shared structure: evidence about one option can update inferences about other options' expected utilities. Structured bandit tasks, in which feature representations constrain reward structure, are consequently viewed as a more faithful abstraction of human decision-making problem in naturalistic settings and have been adopted in recent behavioral studies (e.g., Niv et al., 2015; Schulz et al., 2019; Wu et al., 2018).

From this perspective, two classes of generalization are especially important. The first is within-structure generalization: under a fixed reward-generating structure, experience with one option supports prediction of rewards for related options. The second is meta-generalization under shifts in reward structure. When individuals move across contexts, such as to a town with a different cultural environment, the mapping from features to reward may change. Even so, prior experience can accelerate subsequent exploration by supplying inductive biases about which feature dimensions are likely to be predictive (e.g., cuisine type) and which are unlikely to be predictive (e.g., the first letter of a restaurant's name). Characterizing such meta-level structure learning is therefore central to understanding human decision-making (Collins & Frank, 2013; Gershman & Niv, 2010), and the structured-bandit abstraction provides a route to uncovering context-dependent biases in learning and choice by identifying the features people rely on.

In this paper, we describe an approach to observe meta-level learning in structured bandit tasks. We focus on the *hypothesis space* as a meta-level structural hypothesis in *multidimensional bandit tasks* with multi-attribute options, and aim to infer the hypothesis space from behavioral data. We apply this approach to human choice data to clarify how human learning relates to updates in meta-level structure.

Human learning of structured bandit tasks

In recent years, bandit tasks with structure have been widely used in computational cognitive science to study human learning and generalization. A standard bandit task is defined, for a number of arms $K \in \mathbb{N}$ ($K \geq 2$), by the set of arms $\mathcal{A} = \{1, 2, \ldots, K\}$ and by a real-valued probability distribution P_a (the reward distribution) associated with each arm $a \in \mathcal{A}$. At each time step $t = 1, 2, \ldots, T$, an agent selects one arm a_t^* and receives a stochastic reward $r_t \sim P_{a_t^*}$. The objective is to maximize the (expected) cumulative reward $\sum_{t=1}^{T} r_t$.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

By contrast, in "structured" bandit tasks commonly used in computational cognitive science, each arm is associated with features (Leong et al., 2017; Niv et al., 2015; Schulz et al., 2020). Specifically, there exist a feature space $\mathcal{F}$ and arms $\mathcal{A}$ such that, at each time step $t = 1, 2, \ldots, T$, the agent observes a set of available arms (candidate actions) $a_{1,t}, \ldots, a_{K,t} \in \mathcal{A}$ along with their corresponding features $f_{1,t}, \ldots, f_{K,t} \in \mathcal{F}$. Based on these observations, the agent selects one arm a_t^* (with corresponding feature f_t^*) and receives a reward $r_t \sim P_{f_t^*}$ that is stochastically determined as a function of the feature. The objective remains to maximize the (expected) cumulative reward $\sum_{t=1}^{T} r_t$.

In structured bandit tasks, a wide range of reward structures can be considered depending on how the feature space $\mathcal{F}$ is defined. One line of work examines how humans make sequential decisions based on continuous dimensions (Schulz et al., 2020; Wu et al., 2018). Another line investigates how humans make decisions efficiently when the feature space is high-dimensional and complex (Leong et al., 2017; Niv et al., 2015; Song et al., 2022).

A commonly used computational account of human behavior in bandit tasks is reinforcement learning (RL) based on reward prediction errors. For example, in the K-armed bandit, when the agent selects arm a_t^* at time t and receives reward r_t, the value of the chosen arm $V_t(a)$ is updated as

$$V_t(a_t^*) = V_{t-1}(a_t^*) + \eta(r_t - V_{t-1}(a_t^*)) \tag{1}$$

where $\eta \in (0, 1)$ is a learning rate parameter. Given the resulting values $V_t(a)$, the probability of choosing arm a is typically modeled via a softmax function,

$$\frac{\exp(\beta V_t(a))}{\sum_{a'} \exp(\beta V_t(a'))} \tag{2}$$

where $\beta > 0$ is an inverse-temperature parameter. Models that exploit reward prediction errors have been supported by neuroscientific evidence (Schultz, 1998). However, in the presence of structure, these models are not straightforward to apply. If $\mathcal{F}$ is finite, one could in principle compute a value $V_t(f)$ for every feature $f \in \mathcal{F}$. Yet when $\mathcal{F}$ is high-dimensional, so that the number of possible feature combinations is extremely large, this becomes computationally inefficient. Moreover, when $\mathcal{F}$ is infinite (i.e., continuous), maintaining values for all features is impossible.

In response to these challenges, a variety of models have been proposed in recent years to explain human behavior in structured bandit tasks. In settings with continuous features, models that explicitly learn the mapping from features to expected reward have been developed, and Gaussian process regression with an RBF kernel has been supported as a plausible account (Schulz et al., 2020; Stojić et al., 2020). In contrast, for discrete, multidimensional feature spaces, proposed models include dimension-factorized reinforcement-learning approaches that reduce computational burden (see the next section) as well as accounts based on repeated hypothesis testing (Niv et al., 2015; Song et al., 2022).

A common assumption across these models is that the candidate set of reward-relevant features is known in advance. In the restaurant example, however, successful learning requires acquiring an abstraction of the task structure itself, namely identifying which features are reward-relevant. However, the reward-driven acquisition of meta-level structural knowledge remains a key unresolved issue in computational studies. Here, we propose an analysis framework for the learning of meta-level structure using a multidimensional bandit task. Concretely, we perform inverse inference of the hypothesis space grounded in existing computational models, enabling estimation of the structure participants assume during learning. We then apply the approach to experimental choice data to characterize participants' hypothesis spaces and their transition.

Task and models

Multidimensional bandit task

Niv et al. (2015) investigated human behavior using a multidimensional bandit paradigm. In a multidimensional bandit task, the feature space $\mathcal{F}$ is discrete and factored across multiple dimensions, such that $\mathcal{F}$ is given by the Cartesian product of finite sets $D_1, D_2, \ldots, D_K$: $\mathcal{F} = D_1 \times D_2 \times \cdots \times D_K$.

They used a three-dimensional task with dimensions corresponding to color, shape, and texture:

$$\begin{aligned}\mathcal{F} =& D_{\text{color}} \times D_{\text{shape}} \times D_{\text{texture}} \\ =& \{\text{red, blue, green}\} \times \{\text{circle, triangle, square}\} \\ & \times \{\text{dot, stripe, honeycomb}\}\end{aligned} \tag{3}$$

Rewards were binary (0 or 1) and depended on exactly one of the feature dimensions; participants were told that one of the three dimensions governed reward. For instance, in one condition, only the color dimension affected the probability of reward, with one level (e.g., red) assigned a higher reward probability than the remaining levels (e.g., blue and green).

Models for multidimensional bandit task

Niv et al. (2015) repeatedly administered a multidimensional bandit task while varying which dimension was reward-relevant, and evaluated which behavioral model best predicts participants' choice data. They found that a model termed the *feature RL model with decay* provided the best prediction. In this paper, we refer to this model simply as the *fRL model* and describe it below. Let the value of arm a at time t, $V_t(a)$, be decomposed along its feature vector $f = (f_1, f_2, f_3) \in \mathcal{F}$ into dimension-specific values $W_t(\cdot)$, such that

$$V_t(a) = W_t(f_1) + W_t(f_2) + W_t(f_3) \tag{4}$$

Given the chosen arm a_t^* with features $f_t^* = (f_1^*, f_2^*, f_3^*)$, the model updates

$$W_{t+1}(f_i^*) = W_t(f_i^*) + \eta(r_t - W_t(f_i^*)) \quad (i = 1, 2, 3) \tag{5}$$

and applies decay to features not present in a_t^*,

$$W_{t+1}(f) = (1 - d)W_t(f) \quad (d \in (0, 1)) \tag{6}$$

Choices are then generated by applying the softmax rule (equation (2)) to the resulting arm values $V_t(\cdot)$.

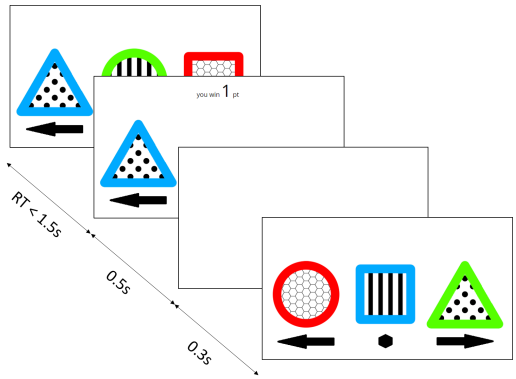

Figure 1: Schematic of the task procedure. On each trial, participants choose one stimulus defined by multiple features and receive points.

In addition to the fRL model, we also consider two comparison models: a classical but computationally inefficient naive RL model, and a Bayesian inference model representing a fully rational inferential strategy. Due to space constraints, we refer the reader to Niv et al. (2015) for details.

Outline of internal hypothesis space estimation

The models described above characterize behavior only under the assumption that the reward-relevant dimensions are already known. In much prior work including Niv et al. (2015), where participants were instructed that color, shape, and texture were relevant: the relevant dimensions are either explicitly provided or otherwise obvious. In real-world decision-making, however, the relevant dimensions are rarely apparent from the outset, and learning these dimensions (i.e., meta-level learning) is therefore important. In such cases, it becomes necessary to recover participants' internal representations of the task structure.

The models described above, and the fRL model in particular, depend on the feature space $\mathcal{F}$. Conversely, replacing $\mathcal{F}$ yields a different feature-based behavioral model. For example, by defining $\mathcal{F} = D_{\text{color}} \times D_{\text{shape}}$ and constructing an fRL model on this space, one obtains a model in which choices are guided only by color and shape. In this paper, we refer to such representations of task structure (i.e., what participants assume could be relevant) as the hypothesis space. By constructing models for all candidate hypothesis spaces and selecting the one that best predicts each participant's behavioral data, we can infer, from a participant's choice sequence, the internal representation of the hypothesis space that they use in sequential decision-making.

Method of experiment

The experimental protocols in this study were approved by the ethics committee of the first author's institution. The ethics review was performed in accordance with the principles of the Declaration of Helsinki.

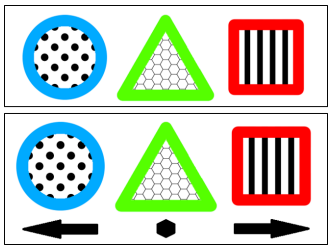

Figure 2: Examples of task stimuli. Stimuli for the basic condition are shown at the top, and stimuli for the position-salient condition are shown at the bottom.

Participants

A total of 61 participants (43 men and 18 women) aged 21-61 years ($M = 40.7, SD = 8.9$) without any color vision deficiencies were recruited through a crowdsourcing service to take part in the online experiment conducted via a web browser. Participants were randomly assigned to either the *basic condition* or the *position-salient condition*. One male participant who did not identify any factors related to the point at the end of the task (described later) was excluded from the analysis for failing to follow the instructions. As a result, 36 participants remained in the basic condition, and 24 in the position-salient condition.

Task

The task was constructed with reference to Niv et al. (2015). On each trial, three stimuli were presented (Figure 1). Each stimulus had three feature dimensions: color (red, blue, or green), shape (square, triangle, or circle), and texture (dot, stripe, or honeycomb). In addition to these dimensions, the spatial position of each stimulus on the screen also affected the probability of receiving points. This manipulation introduced a feature dimension that is likely harder to learn than the visually salient dimensions used in prior work. Accordingly, the feature space was defined as

$$\begin{aligned}\mathcal{F} =& \{\text{red, blue, green}\} \times \{\text{circle, triangle, square}\} \\ & \times \{\text{dot, stripe, honeycomb}\} \times \{\text{left, center, right}\}. \quad (7)\end{aligned}$$

To facilitate learning of the potentially less salient position dimension, we included two experimental conditions. In the basic condition, stimuli were presented as shown in the upper panel of Figure 2. In the position-salient condition, an additional marker indicating position was added beneath each stimulus (lower panel of Figure 2). In both conditions, the simultaneously presented stimuli were constrained to have distinct feature combinations. In both conditions, shape, color, and texture were visually explicit, whereas in the basic condition position was not accompanied by an additional visual cue. Thus, attention to the position dimension was expected to be reduced in the basic condition, while the position-salient condition was expected to increase attention to position relative to the basic condition. We expected participants' learning and the inferred hypothesis space to differ across conditions.

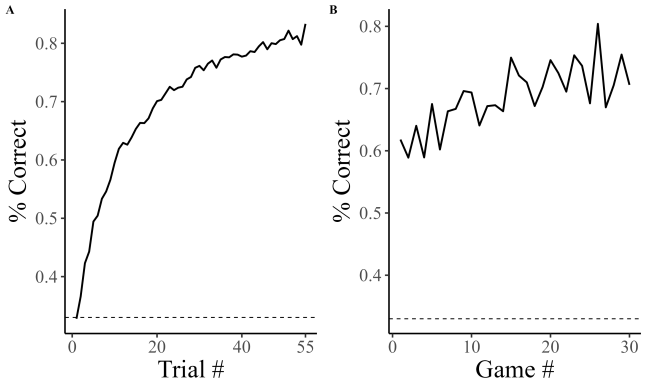

Figure 3: Learning curves showing changes in accuracy. (A) Accuracy at each trial within games. The dashed line indicates chance level ($p = 0.33$). (B) Accuracy for each game.

Upon selecting one of the three stimuli, the two unchosen stimuli were removed and binary feedback (1 or 0 points) was displayed (Figure 1). Participants were instructed to maximize total points. Feedback was shown for 0.5 s. This sequence constituted a single trial. Following feedback offset, a 0.3 s inter-trial interval preceded the onset of the next trial.

The experiment consisted of 30 games, each comprising 45–55 trials. In each game, whether points were obtained depended exclusively on one of four dimensions: shape, color, texture, or position. Each dimension comprised three feature levels (e.g., square, triangle, and circle for shape). Within the reward-relevant dimension, one feature level was designated as high-reward, producing 1 point with probability 0.90, whereas the remaining two feature levels produced 1 point with probability 0.10. The reward-relevant dimension was constrained to switch between successive games.

Participants were required to make a choice within 1.5 s. Failure to respond within this interval triggered stimulus disappearance and a time-out message. Participants then advanced to the next trial by pressing the space bar.

Procedure

Participants completed the experiment online using their own personal computers. To control stimulus presentation timing, participation was restricted to the Google Chrome browser. Before the experiment began, participants were instructed that they would select one stimulus on each trial with the goal of earning as many points as possible, that they should respond as quickly as they could, and that they would receive no points unless they responded within 1.5 s. They were also told that the probability of receiving points depended on a rule and that this rule would change across games; however, no details of the rule were provided.

To ensure that participants understood the response procedure, a practice phase was administered. The practice phase consisted of seven trials, and points were awarded regardless of which stimulus was chosen. After practice, participants completed 30 games of the task described in the previous section. Across games, the high-reward feature was randomly reassigned such that it always differed from the reward-relevant dimension in the preceding game. After completing the task, participants provided an open-ended response to the question, "What do you think influenced how likely you were to obtain points?"

Table 1: Model comparison based on WAIC.

Model	Parameters	WAIC
Naive RL	η, β	1.58×10^5
Bayesian	β	1.38×10^5
fRL (common η)	η, β, d	9.54×10^4
fRL (dimension-wise η)	$\eta^{\text{color}}, \ldots, \eta^{\text{position}}, \beta, d$	9.29×10^4

Result

Learning check

Before conducting model-based analyses, we assessed whether the intended task structure was learned. We defined accuracy on each trial as whether the selected stimulus contained the reward-relevant feature. Statistical analyses were performed in R (version 4.3.2).

We first calculated accuracy by trial index within each game, averaged across all participants and games. The resulting trial-wise accuracy curve is displayed in Figure 3A. Accuracy increased monotonically across trials, and the Spearman rank correlation between accuracy and trial index was significant ($\rho = .994$, $p < .001$), consistent with within-game identification of the reward-relevant dimension.

We then computed accuracy by game index, averaged across participants. Game-wise accuracy is shown in Figure 3B. The Spearman rank correlation between accuracy and game index was also significant ($\rho = .736$, $p < .001$), indicating improved performance over games and increasing adaptation to the task demands.

Model comparison about learning strategy

To assess whether it was appropriate to analyze the data using the fRL model, we evaluated model predictive performance. Because learning rates may plausibly differ across dimensions, we considered two variants of the fRL model: one in which the learning rate was shared across dimensions, and another in which separate learning rates were estimated for the four dimensions, denoted $\eta^{\text{color}}, \eta^{\text{shape}}, \eta^{\text{texture}}, \eta^{\text{position}}$. For this initial model-comparison analysis, we pooled data across the basic condition and the position-salient condition and used the full data.

We adopted a Bayesian approach and performed posterior inference using the `cmdstanr` package. For each chain, we drew 2000 samples, including 400 warm-up iterations, and ran 4 chains (the same sampling configuration was used in all subsequent analyses). Parameters were estimated hierarchically. Specifically, for each participant i, we assumed that individual-

Table 2: Summary statistics of the posterior distributions of group-level parameters in the fRL model.

Parameter	EAP	SD	95% ETI
δ	0.059	0.273	[-0.471, 0.604]
μ_η^{color}	0.126	0.006	[0.114, 0.139]
μ_η^{shape}	0.120	0.006	[0.108, 0.132]
$\mu_\eta^{\text{texture}}$	0.130	0.007	[0.116, 0.144]
$\mu_\eta^{\text{position}}$	0.088	0.006	[0.076, 0.100]
σ_η	0.0449	0.0024	[0.0405, 0.0498]
μ_β	7.23	0.21	[6.81, 7.64]
σ_β	1.56	0.15	[1.28, 1.87]
μ_d	0.484	0.019	[0.448, 0.523]
σ_d	0.127	0.014	[0.103, 0.156]

level parameters η_i, β_i, and d_i were generated from group-level distributions, i.e., $\eta_i \sim N(\mu_\eta, \sigma_\eta^2)$, $\beta_i \sim N(\mu_\beta, \sigma_\beta^2)$, and $d_i \sim N(\mu_d, \sigma_d^2)$. For the model with dimension-specific learning rates, we used a shared standard deviation across dimensions, e.g., $\eta_i^{\text{color}} \sim N(\mu_\eta^{\text{color}}, \sigma_\eta^2)$. We specified priors as $\mu_\eta, \mu_d \sim \text{U}(0, 1)$, $\sigma_\eta, \sigma_d \sim \text{U}(0, 1)$, $\mu_\beta \sim \Gamma(2, 3)$, and $\sigma_\beta \sim \text{U}(0, 10)$. The prior for μ_β followed Niv et al. (2015), whereas the remaining priors were chosen to be uniform distributions that covered a sufficiently broad range of plausible values.

The Gelman–Rubin convergence diagnostic $\hat{R}$ (Gelman et al., 2013) satisfied $\hat{R} < 1.1$ for all parameters in all models, suggesting adequate convergence. We compared models using WAIC, an estimate of expected out-of-sample predictive accuracy (with a complexity penalty); lower WAIC indicates better expected predictive performance (Vehtari et al., 2017; Watanabe, 2013). WAIC values computed from the fitted models are reported in Table 1. As shown in Table 1, the fRL model with dimension-wise learning rates provided the best predictive performance. These results suggest that using fRL for subsequent analyses is comparatively well justified and that it is important to allow for differences in learning rates across dimensions.

Parameter estimation of fRL model

Next, to interpret the parameter estimates obtained from the fRL model with dimension-wise learning rates and to assess whether it is necessary to assume condition differences, we modified the model described in the previous subsection and re-estimated it. Specifically, we modeled the learning rate for the position dimension, η^{position}, as condition-dependent. In each condition, we assumed

$$\eta_i^{\text{position}} \sim \begin{cases} N(\mu_\eta^{\text{position}} - c/2, \sigma_\eta^2) & \text{(basic)} \\ N(\mu_\eta^{\text{position}} + c/2, \sigma_\eta^2) & \text{(position-salient)} \end{cases}. \tag{8}$$

We parameterized the condition difference as $c = \delta\, \sigma_{\eta^{\text{position}}}$ and estimated the standardized difference δ. The prior for δ was set to Cauchy$(0, 1)$.

All parameters satisfied the convergence criterion $\hat{R} < 1.1$, indicating adequate convergence. The posterior summaries for the main parameters are reported in Table 2 (EAP denotes the posterior mean; 95% ETI denotes the equal-tailed interval). The posterior distribution of δ was centered near zero. As a standardized difference, its EAP is extremely small (Cohen, 1988). These results suggest that the learning rate for the position dimension did not differ between the two conditions. Accordingly, in subsequent analyses we pooled the basic and position-salient conditions and analyzed the combined dataset.

In addition, comparing the estimated group-level means of the learning rates across dimensions revealed that $\mu_\eta^{\text{position}}$ was substantially smaller than μ_η^{color}, μ_η^{shape}, and $\mu_\eta^{\text{texture}}$. Indeed, the 95% ETI for $\mu_\eta^{\text{position}}$ showed little overlap with the corresponding 95% ETIs for μ_η^{color}, μ_η^{shape}, and $\mu_\eta^{\text{texture}}$. This pattern suggests that the position dimension was relatively more difficult to learn than the other dimensions.

Table 3: Estimated hypothesis space across participants.

P	C, P	C, S, T	C, S, P	C, T, P	S, T, P	C, S, T, P
1	1	17	2	2	1	36

Note. C: color, S: shape, T: texture, P: position

Estimation of the hypothesis space

Next, to infer participants' hypothesis spaces, we constructed fRL models corresponding to all 15 ($= 2^4 - 1$) candidate hypothesis spaces defined by selecting at least one dimension from the set {color, shape, texture, position}. We fit these models to each participant's data and evaluated prediction performance using WAIC. For each participant, we identified the hypothesis space associated with the model that achieved the minimum WAIC. The result is shown in Table 3.

For 36 out of 60 participants (60%), the full hypothesis space including all four dimensions was supported. This indicates that a majority of participants were able to acquire the correct task structure. For 17 participants (28%), the supported hypothesis space included all dimensions except position. This pattern is consistent with the reduced learnability of the position dimension and suggests that these participants made choices without attending to position.

Hypothesis space transition

To examine how individuals' hypothesis spaces changed over time, we partitioned the 30 games for each participant into several blocks and inferred the hypothesis space separately for each participant and block. Estimating the hypothesis space at the level of a single game failed to converge due to insufficient data; therefore, we grouped five games into one block and analyzed changes across six blocks in total.

For each participant and each block, we identified the hypothesis space corresponding to the model with the lowest WAIC, and computed the number of dimensions included in that hypothesis space (i.e., its breadth). We then averaged this

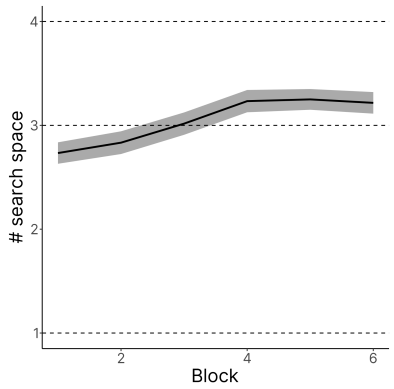

Figure 4: Changes in the breadth of the hypothesis space.

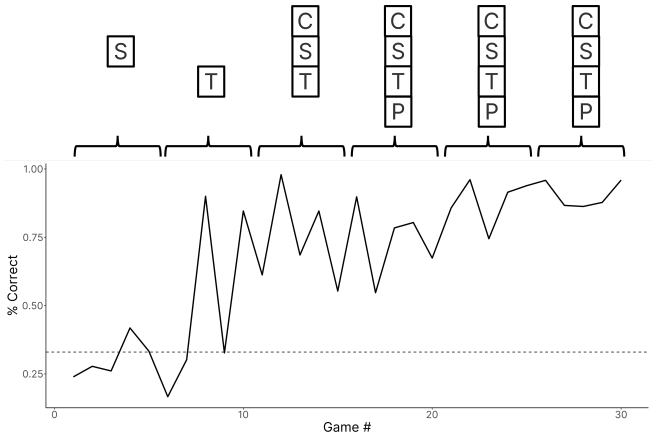

Figure 5: Learning curve and estimated hypothesis space for an illustrative participant.

breadth across participants within each block and visualized its temporal change (Figure 4). The mean hypothesis space breadth was 2.73 in Block 1 and 3.22 in Block 6; moreover, in the later blocks many participants were classified as having the full hypothesis space of breadth 4. The Spearman rank correlation between hypothesis-space breadth and block index was significant ($\rho = .241$, $p < .001$). Thus, in terms of hypothesis space, the results indicate that meta-level task structure was learned over the course of the experiment.

This pattern suggests that learning of meta-level task structure contributed to game-wise learning. As an illustration, Figure 5 jointly shows one participant's learning curve and inferred hypothesis-space trajectory. This participant was selected because both the game-wise improvement and the transition in the inferred hypothesis space were especially clear. In the first two blocks, the participant appeared to have little knowledge of the task structure and made choices under a very narrow hypothesis space. From around Block 3, accuracy increased sharply, accompanied by a shift toward choices consistent with a broader hypothesis space.

Discussion

The goal of this study was to propose a method for inferring the hypothesis space as a task representation in a multidimensional bandit task, and to use this analysis framework to clarify the importance of updating meta-level structure in human learning.

When participants performed the multidimensional bandit task without explicit instruction about the reward structure, the resulting learning curves suggested learning not only within games but also across games. Parameter estimation with the fRL model further indicated, via the pattern of inferred learning rates, that the position dimension was learned more slowly than the visually salient dimensions (color, shape, and texture). This asymmetry suggests that even dimensions that are formally part of the task's generative structure are not necessarily updated uniformly, potentially due to factors such as attentional allocation.

A central contribution of the present work is the proposal of a framework that estimates participants' latent hypothesis spaces by performing inverse inference using behavioral models. Applying this inverse-inference approach, we found that behavior was best explained by the full hypothesis space for 60% of participants, whereas for 28% it was best explained by a hypothesis space excluding position. This result suggests not merely that the position dimension was difficult to learn, but that for some participants it was effectively dissociated from the task representation.

Finally, analyses of transitions in hypothesis space revealed that participants' hypothesis spaces expanded on average as they progressed through the games. This pattern suggests that moving the hypothesis space toward a task-appropriate "true" representation may be associated with improved between-game performance.

The present study has two primary applications. The first concerns basic research on modeling how humans revise their exploration spaces over time. Although we did not directly model the process itself in this study, visualizing internal representations can serve as a tool for generating useful hypotheses about the mechanisms underlying representational transitions. Future work should therefore further analyze and formally model the dynamics of the hypothesis space.

The second application concerns real-world decision-making. By applying the proposed framework to naturalistic tasks, it becomes possible to extract which factors people rely on when making decisions in a given context. In the present study, the position dimension was found to be relatively difficult to learn; this can be interpreted as a bias in meta-level structure learning. Applying analogous methods to choice processes such as restaurant selection or music discovery could enable identification of biases in human structure learning that are grounded in real-world decision problems (Schulz et al., 2019; Wu et al., 2018).

Acknowledgments

This work was supported by JST CREST JPMJCR19A1 and JSPS KAKENHI Grant Numbers 18H03501, 23KJ0793. The authors would also like to thank Yuki Aso for his help in preparing the experimental stimuli used in this study.

References

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2-3), 235–256. https://doi.org/10.1023/a:1013689704352

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman; Hall/CRC.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256. https://doi.org/10.1016/j.conb.2010.02.008

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press. https://doi.org/10.1017/9781108571401

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451–463. https://doi.org/10.1016/j.neuron.2016.12.040

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27. https://doi.org/10.1152/jn.1998.80.1.1

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(28), 13903–13908. https://doi.org/10.1073/pnas.1821028116

Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, *119*, 101261. https://doi.org/10.1016/j.cogpsych.2019.101261

Song, M., Baah, P. A., Cai, M. B., & Niv, Y. (2022). Humans combine value learning and hypothesis testing strategically in multi-dimensional probabilistic reward learning. *PLoS Computational Biology*, *18*(11), e1010699. https://doi.org/10.1371/journal.pcbi.1010699

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179. https://doi.org/10.1016/j.jmp.2008.11.002

Stojić, H., Schulz, E., P Analytis, P., & Speekenbrink, M. (2020). It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, *149*(10), 1878–1907. https://doi.org/10.1037/xge0000749

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*. https://doi.org/10.5555/2567709.2502609

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10.1038/s41562-018-0467-4