OUKA
The University of Osaka
Institutional Knowledge Archive

| | |
|---|---|
| Title | Dynamics of visual attention in exploration and exploitation for reward-guided adjustment tasks |
| Author(s) | Higashi, Hiroshi |
| Citation | Consciousness and Cognition. 2024, 123, p. 103724 |
| Version Type | VoR |
| URL | https://hdl.handle.net/11094/98165 |
| rights | This article is licensed under a Creative Commons Attribution 4.0 International License. |
| Note | |

*The University of Osaka Institutional Knowledge Archive : OUKA*

https://ir.library.osaka-u.ac.jp/

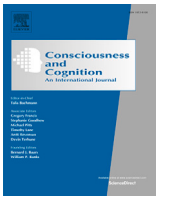
The University of Osaka

Consciousness and Cognition 123 (2024) 103724

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/yccog



Full Length Article

# Dynamics of visual attention in exploration and exploitation for reward-guided adjustment tasks

Hiroshi Higashi

*Graduate School of Engineering, Osaka University, Suita, Osaka, Japan*

ARTICLE INFO

Dataset link: https://github.com/hgshrs/rl_serial

*Keywords:*
Attention
Decision making
Reinforcement learning
Repetition priming
Serial dependence

ABSTRACT

The learning process encompasses exploration and exploitation phases. While reinforcement learning models have revealed functional and neuroscientific distinctions between these phases, knowledge regarding how they affect visual attention while observing the external environment is limited. This study sought to elucidate the interplay between these learning phases and visual attention allocation using visual adjustment tasks combined with a two-armed bandit problem tailored to detect serial effects only when attention is dispersed across both arms. Per our findings, human participants exhibited a distinct serial effect only during the exploration phase, suggesting enhanced attention to the visual stimulus associated with the non-target arm. Remarkably, although rewards did not motivate attention dispersion in our task, during the exploration phase, individuals engaged in active observation and searched for targets to observe. This behavior highlights a unique information-seeking process in exploration that is distinct from exploitation.

## 1. Introduction

Animals acquire knowledge by learning the associations among sensory stimuli, actions, and rewards, thereby enabling better decision-making (Peirce et al., 2019; Rescorla & Wagner, 1972). Reinforcement learning (RL) algorithms (Sutton & Barto, 1998) have been used successfully to explain a broad spectrum of learning behaviors (Daw et al., 2011). These algorithms encapsulate the process of trial-and-error learning, whereby the environment and learner interact through observation, action, and reward. The learning process formulated using the RL algorithm progresses as follows. First, the learner observes the state of the environment and predicts its value (observation stage). Subsequently, based on this predicted value, the learner chooses an action (decision-making stage). In response, the environment delivers a reward (outcome stage). This cycle of observation, decision-making, and outcome is repeated, thus enabling learners to maximize their rewards and learn the optimal actions for a given environment.

Meanwhile, the learning process is understood to encompass two phases: exploration and exploitation (Cohen et al., 2007; Mehlhorn et al., 2015; Walker et al., 2019). During the exploration phase, the learner selects novel actions without certainty as to whether they will yield a reward. This phase enables learners to better comprehend the environment and identify potentially rewarding actions. By contrast, the exploitation phase leverages the information gathered during the exploration phase. In this phase, the learner is confident that a specific action will provide a sufficient reward and thus continues to choose said action. RL-based learners oscillate between the exploration and exploitation phases to adapt to changing environments and maximize their rewards. Consider the following fishing metaphor: Imagine that you attempt to catch fish in various locations. This is an exploratory behavior. Once you discover a place where you can catch numerous fish, you stay there, exploiting this location to the fullest extent possible.

*E-mail address:* higashi@comm.eng.osaka-u.ac.jp.

https://doi.org/10.1016/j.concog.2024.103724
Received 5 March 2024; Received in revised form 24 June 2024; Accepted 26 June 2024

Available online 11 July 2024
1053-8100/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

However, if the yield decreases after some time, you leave the location and commence the search for a new spot. This is a return to the exploration phase.

Exploration and exploitation are believed to utilize distinct cognitive functions during the decision-making stage (Schulz & Gershman, 2019). In the exploration phase, the learner opts for a new action, whereas after transitioning to the exploitation phase, choices tend to be made based on past experiences. Neuroscience research has substantiated these functional and behavioral differences, revealing variations in activation within the insula and dorsal anterior cingulate cortex during the decision-making stage (Daw et al., 2006; Blanchard & Gershman, 2018). Furthermore, the outcome stage presents differences in how values for actions are updated (Findling et al., 2019). Typically, the inconsistency between the expected and the actual rewards is more pronounced in the exploration phase than in the exploitation phase. This discrepancy can lead to distinct neural activity in the ventromedial prefrontal cortex (Blanchard & Gershman, 2018) and orbitofrontal cortices (Summerfield & Koechlin, 2008).

While distinctions in function and neural activity between the exploration and exploitation phases have been elucidated in the decision-making and outcome stages, the observation stage remains unexplored (Walker et al., 2019; Easdale et al., 2019; Walker et al., 2022). According to the definitions of these phases proposed in previous research (Daw et al., 2006), a learner in the exploitation phase knows which information sources in the environment should be observed, focuses their attention on those sources, and ignores the rest (Leong et al., 2017). Moreover, humans exhibit a voluntary attention bias toward information sources that have been valuable in the past, even if these sources are currently task-irrelevant (Anderson et al., 2011; Failing & Theeuwes, 2018; Anderson, 2016). The results for attentional deployment align logically with the definition of the exploitation phase. However, in the exploration phase, the reward does not provide any clues regarding which source should be the focus of attention. In the context of goal-directed learning, how do humans and animals initiate the observation of their surrounding environments? Understanding the process is essential for unraveling the mechanisms underlying our information-seeking ability.

This study investigates the differences in attention allocation between the exploration and exploitation phases. In contrast to the exploitation phase, where the information sources to be observed are known and attention is focused on the target source (Leong et al., 2017), we hypothesized that during the exploration phase, learners would distribute their attention to include observable information sources of which their values are uncertain. We aim to find behavioral evidence supporting this hypothesis. To achieve this goal, we designed experimental tasks that integrated a two-armed bandit problem (Sutton & Barto, 1998), visual adjustment tasks (Pilly & Seitz, 2009), and serial effects (Dong & Atick, 1995; Wiggs & Martin, 1998; Schacter et al., 2004).

In solving a two-armed bandit problem, wherein a learner chooses between two options (arms) to maximize rewards, actions can be categorically defined as either exploratory or exploitative on a trial-by-trial basis. Each arm represented a unique source of information. In our study, human participants were presented with two such arms, represented either as random dot motion (RDM) stimuli in Experiment 1 or as randomly oriented bar stimuli in Experiment 2. The participants were tasked with selecting one of the two arms (the target arm) and reporting on its attribute—either motion direction or tilting orientation. The precision of their responses, quantified by the adjustment error, served as an indicator of the extent of attention paid to the target arm. Furthermore, we analyzed serial effects—specifically, repetition priming (Schacter et al., 2004) in Experiment 1 and serial dependence (Fischer & Whitney, 2014) in Experiment 2—across consecutive trials. An observed increase in the serial effects of the stimuli associated with the non-target arm from the preceding trial would support our hypothesis, demonstrating how attention is allocated between target and non-target sources.

The two experiments employed distinct visual stimuli and experimental forms: Experiment 1 utilized RDMs in an offline, in-face setting, whereas Experiment 2 featured randomly oriented bar stimuli in an online environment. Despite the differing experimental designs, both experiments demonstrated serial effects linked exclusively to the non-target arm exclusively during the exploration phase. These consistent findings across diverse experimental setups support the validity of our hypothesis, demonstrating attentional dispersion during the exploration phase.

## 2. Experiment 1

Experiment 1 was conducted in a more controlled experimental environment than Experiment 2, which was conducted online. The arms in Experiment 1 were represented by two sets of RDMs: one black and one white. We introduced a covert condition between successive trials to potentially trigger repetition priming, which is a well-known cognitive effect whereby a previously encountered stimulus is recognized more quickly and accurately than a novel stimulus (Pinkus & Pantle, 1997; Anstis & Ramachandran, 1987; Laarni, 1999; Campana, 2002).

### *2.1. Material and methods*

#### *2.1.1. Participants*

Using G*Power estimation (Faul et al., 2007), we determined that a minimum of 32 sessions would be required to achieve an effect size $f$ of 0.25 with a statistical power of 0.8 for a two-way analysis of variance (ANOVA) with two factors for each method (see Fig. 3 in Section 2.2). Seven sessions were conducted for each participant. As a session might not include trials with desired states, which were defined by the learner's phase (see Section 2.1.7), we continued adding participants until the required number of sessions was reached. Finally, 18 participants (13 male and 5 female participants) participated in the experiment.

This study was approved by the Committee for Human Research at the Graduate School of Informatics, Kyoto University, and was conducted in accordance with the Declaration of Helsinki. All participants signed an informed consent form before participating and

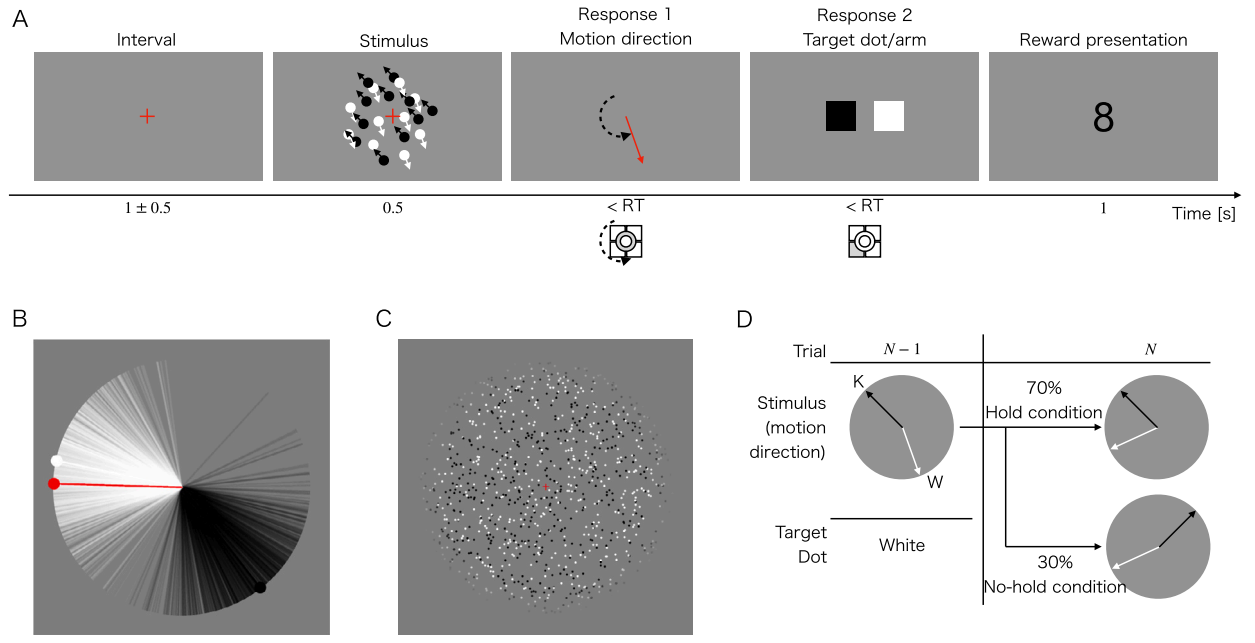



**Fig. 1.** Task procedure. **A**. Flow of a single trial. After an inter-trial interval of $1 \pm 0.5$ s, an RDM stimulus was presented. Participants responded by identifying the average motion direction for the targeted colored dots and then specifying the target color. A reward was subsequently given, determined by the reward function for the target dot color, and the adjustment error of the motion direction average. **B**. Example of motion directions for dots. The motion directions (illustrated by the black and white lines from the center) were randomly distributed (depicted by the black and white dots). The red line represents an instance of a participant's response during a trial. For this trial, the target dot color was white, while the response's direction (red) was close to the average motion direction of the white dots. **C**. A single frame of the RDM stimulus used in our experiment. **D**. Hold and No-hold trial conditions. 70% of all trials were randomly assigned to the Hold condition, with the remaining 30% being assigned to the No-hold condition. For trials in the No-hold condition, the motion direction average for each dot color was randomly determined. Conversely, the Hold condition retained the motion direction average for the dot color, not selected as the target color by participants in the preceding $(N-1)$ trial.

were compensated with a monetary reward. They had normal or corrected-to-normal visual acuity and were aged between 19 and 24 years (mean: $21.3 \pm 1.4$ years).

### *2.1.2. Individual trial design*

The RDM stimuli, composed of black and white dot motions in different directions, were presented to the participants. The participants were asked to indicate the average motion direction of either the black or white dots, with the choice of dot color being based on the individual participant. At the end of each trial, a reward was given. The reward was determined based on a two-armed bandit problem, considering which color (black or white) yielded a higher reward, and the error in estimating the motion direction. That is, the task integrated elements of both learning and perceptual decision-making. A single trial comprised the presentation of the RDM, determination of the motion direction, selection of the target dot color, and delivery of the reward (Fig. 1A). Each participant was seated on a stool as they observed a 23-inch LCD monitor located 60 cm away. The background color was gray (RGB: 128, 128, 128).

An RDM stimulus was presented for 0.5 s, following a red fixation cross (0.41° in size) that appeared $1 \pm 0.5$ s before the RDM presentation. The stimulus was displayed within a circular aperture with a radius of 8.5° at the center of the monitor, and it faded at the edge of the circle. The stimulus dots were black (RGB: 0, 0, 0) and white (RGB: 255, 255, 255). Each dot had a diameter of 0.15°. They moved at a rate of 4.2°/s. The frames for the presentation were refreshed at 60 Hz. An average of 500 dots was displayed for each color. During the presentation of the RDM stimulus, the fixation cross remained on the screen. Participants were instructed to keep their eyes on the fixation cross while it was visible.

After the RDM presentation, the participants were required to make two responses regarding the average motion direction and target dot color. In the first response, participants were asked to identify the average motion direction of either the black or white dots. They controlled a red bar rotating at the center of the monitor using a trackball wheel. Half of the participants submitted their responses by clicking the left button followed by the right button, while the other half clicked the right button followed by the left. For the second response, the participants were asked to specify the color corresponding to the motion direction they had identified in the first response. Half of the participants double-clicked the left button to choose black and the right button to choose white, while the other half double-clicked the right button to choose black and the left to choose white. These two response steps waited until the participant provided both responses. Finally, the reward was presented numerically at the center of the monitor for 1 s.

*2.1.3. Motion direction design and its priming*

The motion direction for each random dot was determined as follows: The average motion directions of the black and white dots were denoted by $\mu_{\mathrm{K}}$ and $\mu_{\mathrm{W}}$, respectively. The motion direction $d$ was sampled from a Von-Mises distribution, $d \sim \mathrm{VM}(\mu_a, 7)$ for $a = \mathrm{K}, \mathrm{W}$, where $\mathrm{VM}(\mu, \kappa)$ represented a Von-Mises distribution with a mean of $\mu$ and a variance of $\kappa$. The initial locations of the dots on the monitor were randomly selected. Fig. 1B illustrates the randomly generated motion directions and the participant's response to the stimulus. An example of a frame of an RDM stimulus is displayed in Fig. 1C.

The repetition priming on the motion direction adjustment was induced by the following hidden condition: The motion direction averages for each trial were determined by two conditions: *Hold* and *No-hold*. In the No-hold condition, the directions, $\mu_a \in [0, 360°)$, for $a \in \{\mathrm{K}, \mathrm{W}\}$, were randomly selected, with the constraint that $D(\mu_{\mathrm{K}}, \mu_{\mathrm{W}}) \geq 60°$, where $D(x, y) > 0$ defines the difference between the two motion directions, $x$ and $y$, as $D(x, y) = \min(|x - y|, |x - y + 360|)$. By contrast, the Hold condition was designed to induce a positive repetition priming effect between consecutive trials, thus creating a connection between the preceding and current trials, as shown in Fig. 1D. This condition preserved the motion direction of one dot color, specifically the non-target dot color, from the preceding $(N - 1)$ trial; that is, the dot color whose motion direction the participant did not choose in the preceding trial. The trials were assigned to the two conditions as follows: 30% to the No-hold condition and 70% to the Hold condition.

*2.1.4. Two-armed bandit problem design*

In this task, participants were asked to choose the color of the dots (either black or white) for which they would respond to the motion direction. These dot colors acted as the "arms" in the two-armed bandit problem, representing the two available options for the participants to select in pursuit of a higher reward. The reward was determined based on the maximum potential reward for the target arm and the error between the actual motion directions and those indicated in their response. The maximum reward for each arm ranged from 1 to 10. In [each trial, the maximum reward for each arm gradually changed by a random value using a random walk strategy such that $\Omega_a = \Omega_a \pm \mathcal{U}(0, 5)$, $a \in \{\mathrm{K}, \mathrm{W}\}$, where $\Omega_{\mathrm{K}}$ and $\Omega_{\mathrm{W}}$ represent the maximum potential rewards for the black and white arms, respectively, and $\mathcal{U}(x, y)$ is a sample from a uniform distribution ranging from $x$ to $y$. Given $s$ as the target arm, the delivered reward was based on $\Omega_s$ and was reduced according to the error between the actual motion direction $\mu_s$ and the response motion direction $\nu$, given as

$$r = \mathrm{round}\left(\Omega_s \frac{f_{\mathrm{VM}}(\nu \mid \mu_s, 5)}{f_{\mathrm{VM}}(0 \mid 0, 5)}\right) \tag{1}$$

where $f_{\mathrm{VM}}(\nu \mid \mu, \kappa)$ is the probability density function at $\nu$ for a Von-Mises distribution with a mean of $\mu$ and a variance of $\kappa$, defining the degree to which the error impacts the reward. The denominator was employed solely for normalization purposes, ensuring that the maximum reward equaled $\Omega_s$. For example, consider a scenario wherein the motion direction of the target arm $\mu_a$ is 60°, the response's motion direction $\nu$ is 85°, and the maximum reward for the target arm is 7. The error in the motion direction is 25° and the reward in this case would be 4.

*2.1.5. Session design*

Each session comprised 50 trials. Participants underwent two practice sessions and seven recording sessions. In the first practice session, with the reward delivery, the average motion directions of the dots were presented to the participants as the orientations of black and white bars, and the response motion direction was shown as the orientation of a red bar. The second practice session did not display these bars. To conceal the existence of the Hold and No-hold conditions, neither of the two practice sessions included a trial of the Hold condition. During the practice sessions, we ensured that the participants understood the distribution of random dot motion directions, the trial-by-trial variations in maximum rewards, and the impact of motion direction errors on reward reduction. Participants were unaware of the existence of the Hold and No-hold conditions according to the post-test questionnaire. Between-session intervals lasted at least one minute.

*2.1.6. Reinforcement learning model*

Our task fused the two-armed bandit problem with the RDM task. To analyze the behavior from the perspective of the bandit problem, we employed a simple reinforcement learning model. The model focused solely on predicting one of two captured responses—specifically, the target arm, rather than the motion direction.

Within the model, two value functions are defined: $Q_{\mathrm{K}}$ and $Q_{\mathrm{W}}$, which represent the expected rewards when choosing the black and white dots as the target arm, respectively. At the start of a session, these value functions were initialized to $\delta$. Given $s \in \{\mathrm{K}, \mathrm{W}\}$ is the participant's response—i.e., the arm for which the participant responded to the motion direction—and $r$ is the delivered reward. Subsequently, the model based on $Q$-learning (Sutton & Barto, 1998) updated the value function $Q_s$ as

$$Q_s = Q_s + \alpha(r - Q_s), \tag{2}$$

where $\alpha$ denotes the learning rate. Conversely, for the non-target arm $s'$, the value function was updated as

$$Q_{s'} = Q_{s'} + \gamma(\delta - Q_{s'}), \tag{3}$$

where $\gamma$ indicates a forgetting parameter that facilitates the convergence of the value function to the initial value $\delta$ when the arm was non-targeted.

The model was evaluated based on the likelihoods associated with the target arm. For an individual trial, the likelihoods, $P(\mathrm{K})$, for the black arm and $P(\mathrm{W})$ for the white arm, were computed as

$$P(a) = \frac{\exp(\beta Q_a)}{\exp(\beta Q_{\mathrm{K}}) + \exp(\beta Q_{\mathrm{W}})}, \tag{4}$$

for $a = \mathrm{K}, \mathrm{W}$, where $\beta$ denotes the inverse temperature parameter. The optimal values for the parameters—$\alpha$, $\gamma$, $\delta$, and $\beta$, which together maximize the cumulative likelihood across a session—were determined using the sequential least squares programming (SLSQP) method. As a starting point for this optimization process, the parameters were initialized at specific values: 0.5 for $\alpha$ and $\gamma$, 5 for $\delta$, and 1 for $\beta$.

For each trial, we identified the participant's learning phase as either exploration or exploitation. We used the calculated likelihood for the target arm, $P(s)$, to categorize the trial into two distinct phases. If the likelihood exceeded 0.75, we classified the trial as part of the exploitation phase. In this phase, there was a notable difference between the estimated value functions for the black and white arms, thereby leading participants to confidently choose the arm with the highest value. Conversely, if the likelihood was less than 0.75, we classified the trial as part of the exploration phase. In this phase, the difference between the value functions was marginal.

#### 2.1.7. Motion direction adjustment error

We assessed the participants' attention to the stimulus by examining the errors in their adjustment of motion direction. Our underlying assumption was that a smaller error would indicate a higher level of attention (Alais et al., 2017). The error, denoted as the motion direction adjustment error (AE), was defined as follows: Given the average motion directions for the black and white dots—represented by $\mu_{\mathrm{K}}$ and $\mu_{\mathrm{W}}$, respectively, the participant's estimated motion direction $v$ (obtained from their first response), and the target arm $s \in \{\mathrm{K}, \mathrm{W}\}$ (obtained from the second response)—AE [°] was computed as

$$|\mathrm{AE}| = D(\mu_s, v) = \min(|\mu_s - v|, |\mu_s - v + 360|). \tag{5}$$

The smaller the AE, the more accurately the participants were able to estimate the motion direction, suggesting greater attention to the stimulus.

#### 2.1.8. State of trial

In our analysis, we only considered error from trials that met specific conditions. To determine whether trial $N$ satisfied these conditions, we applied the following criteria:

1. Change of target arm: The target arm must have changed from trial $N-1$ to trial $N$. This allowed us to investigate a repetition priming effect.
2. Exploration phase: Trial $N$ must have been in the exploration phase. This criterion helped in eliminating outlier trials wherein participants altered the target arm even in the exploitation phase.

After applying these conditions, we classified the remaining trials according to the following conditions:

- Learning phase (Explor/Exploi): According to the RL model fitting, whether a participant was in the exploration (Explor) or exploitation (Exploi) phase during the stimulus presentation in the preceding trial.
- Repetition (Hold/Nohold): Whether the current trial was in the Hold or No-hold condition.

This classification helped structure our analysis, enabling us to assess the behavior of participants in different scenarios.

The learner's phase, whether in the exploration or exploitation phase, was not controllable. Consequently, certain sessions may not have included any trials in the aforementioned states. We also applied the following exclusion criteria for trials and sessions to refine our analysis: First, we removed trials wherein the AE exceeded the 90th-percentile, as these were deemed outliers. Furthermore, we excluded all sessions comprising 10 or more outlier trials. Consequently, for the statistical analysis shown in Fig. 2, 925 samples (14.7%) were excluded, resulting that eight sessions (6.3%) were excluded (no participant exclusion) from the original set. For our state-based analysis shown in Fig. 3, we derived a session-wise AE for each state by averaging the trials classified into a specific state within a session. As we mentioned in Section 2.1.1, because of the possibility that a session might not include trials of certain states, any sessions that did not include trials corresponding to all the states were also excluded. Consequently, 87 sessions (70.6% from the original set), including all sessions from five participants (27.8%), were excluded from this analysis.

### 2.2. Results and discussion

First, we present a statistical analysis of the participants' first response, wherein they responded to the motion direction of the target arm. In our behavioral experiment, the participants responded to the motion direction with an average error of $15.959 \pm 17.585°$ (Fig. 2A). On average, this adjustment error reduced the delivered reward ($r$ computed by (1)) from the maximum reward ($\Omega_{\mathrm{K}}$ or $\Omega_{\mathrm{W}}$) by 17.5%. For instance, if the maximum reward is 7, the delivered reward is reduced to 6. The 90th-percentile is also shown, and trials with an AE greater than this threshold were excluded from further analysis.

Fig. 2B shows the session-wise AEs for each trial condition (Hold/No-hold) and target arm continuity (Keep/Change). The target continuity for a trial was defined as whether the target arm was kept from the preceding trial (Keep) or changed (Change). A four-way ANOVA was performed on 469 samples from 118 sessions involving 18 participants to examine the influence of trial condition, target continuity, their interaction, participant, and session on AE. The test revealed effects of the target continuity ($F(1, 348) = 6.824, p =$

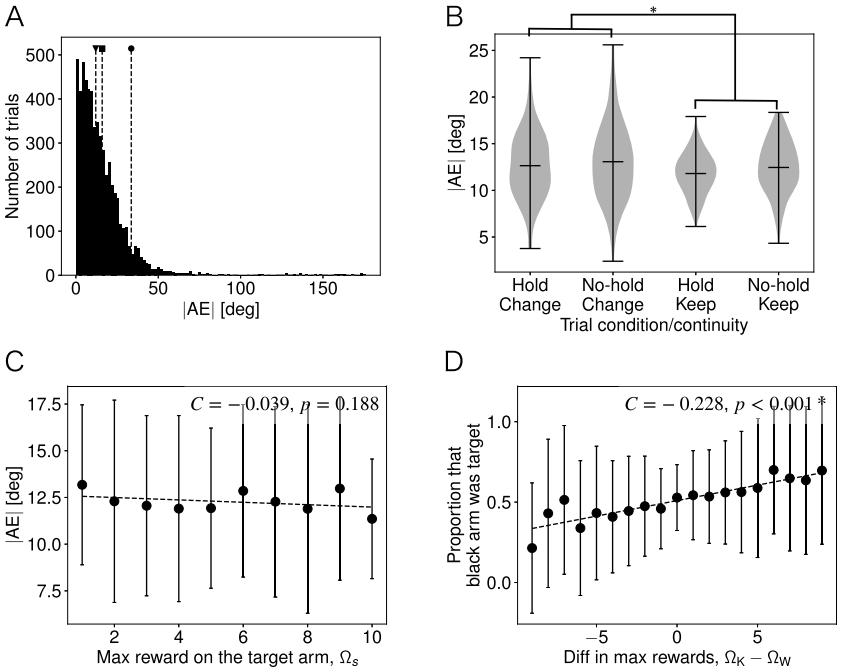



**Fig. 2.** Statistics of behaviors (AE and target arm). **A**. The histogram of AE. The average is denoted by a dashed line with a triangle marker, the median by a dashed line with a square marker, and the 90th-percentile (33.606°) by a dashed line with a circle marker. **B**. AE for the different trial conditions (Hold/No-hold) and the target arm continuity. "Change" refers to the trials wherein the target arms differed from the preceding trial, while "Keep" represents the trials where the target arms remained consistent with the preceding trial. **C**. The relation between the maximum potential rewards on the target arm and the session-wise AEs. For averaging within each session, the maximum rewards for each trial were rounded to the nearest integer. The error bars represent the standard deviation over the sessions. **D**. The relation between the difference in the maximum potential rewards on the two arms, $\Omega_K - \Omega_W$, and the proportion of counts that the black arm was chosen as the target. The reward differences for a trial were rounded to the nearest integer and the proportion at each difference were averaged within a session. The error bars represent the standard deviation over the sessions.

0.009) and participant ($F(17, 348) = 6.824, p < 0.001$). This suggests that the adjustment errors were larger when the participants changed the target arm from the preceding trial. No effect on the interaction between the trial condition and target continuity was found ($F(1, 348) = 0.002, p = 0.966$), indicating that, when not considering the learning phase, the analysis did not detect any repetition priming effect on the adjustment. Fig. 2C provides no evidence ($p > 0.05$) of a correlation between the maximum reward on the target arm and the AE. A Pearson correlation test ($df$: 1,167) revealed a correlation coefficient of −0.039 with a $p$-value of 0.188. This finding suggests that the AE was not significantly influenced by the magnitude of the actual or expected reward. Subsequently, we present an analysis of the second response, wherein the participants indicated which arm they used to adjust the motion direction in the first response. Fig. 2D reveals that the proportion of counts that the participants selected the black arm as the target depended on the difference in the maximum potential rewards, denoted by $\Omega_K - \Omega_W$. When the maximum reward for the black arm was higher than that for the white arm, the participants more frequently chose the black arm. A Pearson correlation test ($df$: 1,505) between the proportion and the reward difference revealed a correlation coefficient of −0.228 ($p < 0.001$). These results suggest that the participants' choices were goal-directed, aiming to maximize their rewards. Their decisions to select a particular target arm were influenced by the relative rewards associated with the black and white arms, which reflected their strategic approach to the task.

We conducted a comparative analysis of the AE across four trial states—Explor-Hold, Exploi-Hold, Explor-Hold, and Exploi-Nohold—to assess the impact of the preceding trial's RL phase (exploration or exploitation) on the AE, as depicted in Fig. 3A. A four-way ANOVA was performed on 152 samples from 38 sessions involving 13 participants. This analysis examined the influence of the preceding trial's phase, condition (Hold or No-hold), interaction between the phase and condition, participant, and session on the AE. The results revealed significant differences attributable to the interaction between phase and condition ($F(1, 111) = 4.599, p = 0.034$) and participants ($F(12, 111) = 2.019, p = 0.028$). Further, a Tukey's honest significant difference test for pairwise comparisons of the four states indicated that the AE for State Explor-Hold was significantly lower than that for State Exploi-Hold ($p = 0.046$) and for State Explor-Nohold ($p = 0.032$). These results demonstrate that the repetition priming effect was more pronounced when a preceding trial was in the exploration phase, as opposed to the exploitation phase. The AEs under the No-hold condition suggest that when no serial connection exists—as in the No-hold condition—the phase in the preceding trial does not influence the motion direction adjustment in the current trial. Furthermore, unlike the exploration phase, the exploitation phase did not exhibit a repetition priming effect, as evidenced by the lack of significant differences between States Exploi-Hold and Exploi-Nohold. Overall, by comparing the four states, we observed that the repetition priming effect occured only when the preceding trial was in the exploration phase. This indicates that, during the exploration phase, the participants paid more attention to the motion direction of dots associated with the non-target arm.

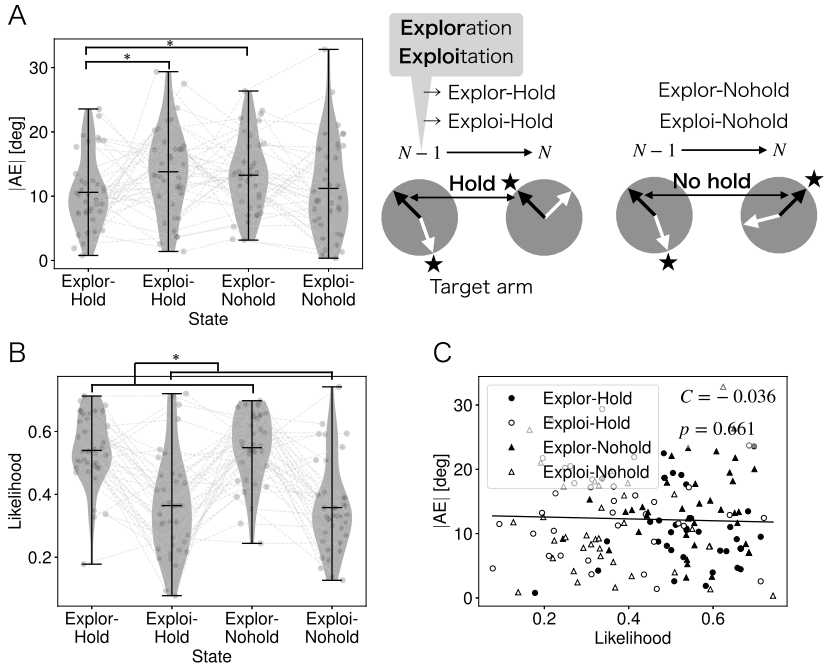



**Fig. 3.** Analysis with the trial state. **A**. AEs for each state. Horizontal lines indicate, from top to bottom, the maximum, average, and minimum values. Gray dots represent AEs for each session, while gray shadows illustrate AEs' probability density. Dashed lines connect the samples from the same session. **B**. Likelihood for the target arm in the current trial for each state. **C**. Likelihood and AE at each state per session. Dots denote individual sessions and the line represents their approximation line.

The results shown in Figs. 3B and 3C suggest that AE was predominantly influenced by repetition priming. As shown in Fig. 3B, a four-way ANOVA indicated that the participant ($F(12, 111) = 3.463, p < 0.001$) and the phase of the preceding trial ($F(1, 111) = 84.971, p < 0.001$) significantly affected the likelihood of choosing the target arm in the current trial. This effect of the preceding phase on the likelihood might be explained by the transition from the exploitation phase to the exploration phase. During this transition, the non-target arm (i.e., the target arm in the preceding trial with a high likelihood in that trial) may still maintain a high likelihood in the current trial, resulting in a relatively lower likelihood for the new target arm in the current trial. Conversely, as shown by a Pearson correlation test with a correlation coefficient of $-0.036$ and a $p$-value of 0.661, there was no significant correlation between the likelihood and AE (3C). While Fig. 3B reveals that the phase in the preceding trial influenced the current likelihood—indicating lower expected rewards for the target arm in the current trial if the preceding trial was in the exploitation phase—Fig. 3C implies that this expectation did not affect AE. Thus, repetition priming may be the primary driver of variations in AE.

## 3. Experiment 2

In Experiment 1, we measured attention allocation for the non-target arm by identifying the serial effect known as the repetition priming effect. Experiment 2 employed a similar concept, but instead focused on another serial effect called serial dependence. Serial dependence is a behavioral bias wherein current visual stimuli and decisions are influenced by previous seen stimuli (Fischer & Whitney, 2014; van Bergen & Jehee, 2019). If we detect the bias from the non-target arm in the preceding trial, this would provide evidence that humans pay attention not only to the target arm but also to the non-target arm.

Experiment 2 was conducted online, presenting a less controllable experimental environment but allowing for the participation of numerous number of observers (over 200). In this experiment, the arms were symbolized by two sets of bars, one black and one white, each with orientations determined by a specific distribution. Participants asked to adjust the average orientation of the bars corresponding to either black or white arm. We investigated whether serial dependence from the preceding trial influenced the orientation adjustment in the current trial.

### 3.1. Material and methods

#### 3.1.1. Participants

The number of participants was based on a precedent online experiment related to serial dependence (Xia et al., 2016). We recruited 213 participants via Prolific. All participants provided their informed consent online. This study adheres to the procedures approved by the Committee for Human Research at the Graduate School of Engineering, Osaka University, and compiled with the Declaration of Helsinki. The participants received a compensation of £6 for completing the entire experiment, with a chance to

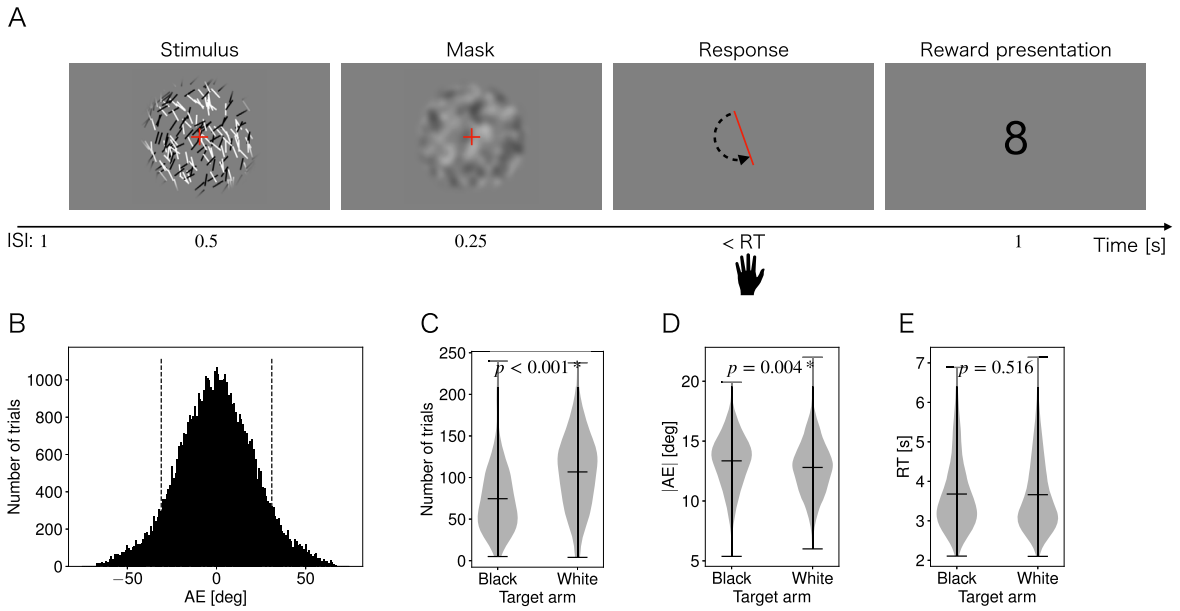



**Fig. 4.** Behavioral results on the orientation adjustment task. **A**. Overview of the task. **B**. Histogram of the adjustment error (AE). The dashed line marked the 90th-percentile (31°) of AE threshold for sample exclusion. **C**. Counts of selections for the black and white arms. **D**. Distribution of adjustment errors for the black and white arms. **E**. RTs for adjustments on the black and white arms.

earn a performance-based bonus ranging from £0 to £4 (average: £2), determined by the rewards earned in our task. We conducted participant and sample screening (for details, see Section 3.1.5), resulting in a final sample of 180 participants.

### *3.1.2. Individual trial design*

The participants were asked to perform an orientation adjustment task featuring bars in black and white bars, each set at different orientations. They were instructed to estimate the average orientation of either black or white bars, with the choice of color left to their discretion. At the end of the trial, participants received a reward determined by a two-armed bandit problem. This required them to select the color of bars they had estimated, considering which color had historically offered the highest rewards. Additionally, the reward amount was adjusted based on the error in their orientation estimation. A visual overview of this task is presented in Fig. 4.

A single trial commenced with the presentation of a red fixation cross against a gray background (RGB: 128, 128, 128), which remained visible for 1 s before the onset of the bar stimulus. Following the fixation period, a stimulus comprising spatially random located black and white bars, was displayed for 0.5 s. Each color set comprised approximately 100 bars, all of which varied in orientation. The mean orientations of the black and white bars, $\mu_{\mathrm{K}}$ and $\mu_{\mathrm{W}}$, respectively, were quasi-randomly assigned from the range of $[0, 180°)$, ensuring a minimum difference of 30° between $\mu_{\mathrm{K}}$ and $\mu_{\mathrm{W}}$. The orientation of each bar was randomly determined within a uniform distribution of $U(\mu_a - 30°, \mu_a + 30°)$ for $a = \mathrm{K}, \mathrm{W}$, where $U(x, y)$ denotes a uniform distribution bounded by $x$ and $y$. The bars were placed at random locations in a circle at the monitor's center.

After the bar stimuli were displayed, the participants were asked to estimate the average orientation of either the black or white bars. The participants used the left and right cursor keys on their keyboard to adjust the orientation of the red bar displayed at the center of the monitor and finalized their answers by pressing the space key. The response period lasted until an answer was provided. After the response, the reward was presented numerically at the center of the monitor for 1 s.

### *3.1.3. Two-armed bandit problem design*

In this task, the participants were instructed to select the color of the bars (either black or white), whose orientation they estimated. These color choices functioned as the "arms"' in the two-armed bandit problem, offering two options for participants to choose from in reward-guided decision-making. The outcome reward's determination was contingent upon the maximum potential reward of the chosen "arm" and the discrepancy between the actual orientation and the participant's estimation. The maximum potential reward for each arm ranged from 1 to 10. In each trial, the maximum reward for each arm was adjusted using a random walk strategy, denoted by $\Omega_a = \Omega_a \pm p$ for $a = \mathrm{K}, \mathrm{W}$, where $\Omega_{\mathrm{K}}$ and $\Omega_{\mathrm{W}}$ denote the maximum potential rewards for the black and white arms, respectively. The variable $p$ took the value of 0 or 2 with an equal likelihood.

The outcome was determined by the maximum potential reward and adjustment error (AE) in orientation. Let $v$ represent the orientation indicated in participants' responses. The adjustment error of the target arm was defined as

$$\mathrm{AE} = D(\mu_{\hat{a}}, v) = \begin{cases} \mu_{\hat{a}} - v - 180 & (\mu_{\hat{a}} - v > 90) \\ \mu_{\hat{a}} - v + 180 & (\mu_{\hat{a}} - v < -90) \\ \mu_{\hat{a}} - v & (\text{otherwise}) \end{cases}, \tag{6}$$

where $\hat{a}$ is the target arm, estimated by $\hat{a} = \arg\min_{a\in\{\mathrm{K,W}\}} D(\mu_a, v)$. The outcome reward, $r$, was calculated by

$$r = \mathrm{round}(\min(\Omega_{\hat{a}}(30 - |\mathrm{AE}|)/30, 0)). \tag{7}$$

If the AE was 0, the outcome remained the maximum potential reward. Conversely, if the AE exceeded 30°, no reward was provided.

### 3.1.4. Session design

Each participant partook in multiple sessions: One practice session focused on a basic orientation adjustment task, while another focused on a simple two-armed bandit task; another focused on a reward-guided orientation adjustment task (our specified task), which was followed by the main recording session. The entire experimental procedure lasted approximately 45 min.

In the practice session for the simple orientation adjustment task, participants were presented with a stimulus, as illustrated in Section 3.1.2, with the difference being that each trial contained bars of only a single color. The color of each trial, black or white, was selected randomly. The participants were asked to estimate the average orientation of the bars. Following their response, both the correct and reported orientations were displayed, represented by bars in black or white for the correct orientation and red for the participant's reported orientation. This process was repeated across 10 trials.

In the second practice session, the participants were acquainted with the two-armed bandit problem, which involved the selection of two options, represented by black and white squares. Following their choice, the participants received a reward corresponding to the selected arm. The reward value for each arm varied over time, determined as outlined in Section 3.1.3. The participants aimed to maximize their total reward across 25 trials, each comprising arm selection and reward presentation stages.

The final practice session introduced participants to the reward-guided orientation adjustment task. This session aimed to help the participants understand how adjustment errors could affect the rewards. The task followed the procedures detailed in Sections 3.1.2 and 3.1.3, with the key addition that orientation feedback was provided to the participants. Along with the outcome, the averaged orientations of the black and white bars, as well as the participants' responses, were depicted using black, white, and red bars, respectively. Participants were instructed to maximize their rewards by accurately estimating the arm with the highest potential reward (the target arm) and precisely adjusting the orientation of the target arm across 50 trials.

The recording session was divided into four blocks. In each block, participants engaged in the orientation adjustment task 50 times without receiving orientation feedback. Only the outcome reward was presented as feedback for each trial and, the objective was to maximize the rewards within each block. Upon completing a block, the potential maximum rewards were reset and the participants were given a break lasting at least one minute.

### 3.1.5. Data screening

We applied specific exclusion criteria to both the samples and participants. Initially, 213 participants contributed 42,600 samples (50 trials × 4 blocks × 213 participants). By applying the following criteria, we retained 32,451 samples from 179 participants.

Trial samples with an adjustment error magnitude |AE| exceeding the 90th-percentile and reaction times (RTs) for the adjustment outside the range of 1.5 s to 10 s were excluded, resulting in the omission of 17,547 samples (32.9%). Furthermore, participants were excluded if more than 80% of their trials were rejected or if their average reward was below 1.75. This threshold for the average reward was determined based on the results from 10,000 runs of random responses of 1.71. Additionally, we excluded participants who, after labeling the exploration and exploitation phases (details in Section 3.1.6) for each trial sample, lacked either phase across all their samples. In total, 34 participants (15.9%) were excluded from the study.

### 3.1.6. Reinforcement learning model

We employed the reinforcement learning outlined in Section 2.1.6. To label each trial sample in either the exploration or exploitation phase, we established a threshold: a likelihood of 0.6 for choosing the target arm. To determine the optimal values for the RL parameters, $\alpha$, $\gamma$, $\delta$, and $\beta$, which together maximized the cumulative likelihood across blocks for each participant, were determined using SLSQP.

### 3.1.7. Analysis for serial dependence

The serial dependence effect is a bias in current decision influenced by previously seen stimuli (Fischer & Whitney, 2014; van Bergen & Jehee, 2019). In this experiment, the adjustment of bar orientation could be biased by the orientations seen in the preceding trial. Specifically, the adjustment might shift towards the orientations in the preceding trial, causing a bias in the adjustment error, which is the difference between the current stimulus orientation and the adjustment. Our stimuli included two orientations corresponding to the black and white bar sets, one of which was the target and the other the non-target. We investigated whether this bias was present for both the target-arm ΔT and the non-target arm ΔNT.

The magnitude of the serial dependence was quantified by fitting to the adjustment error (response variable) from Δ, which represents the difference between the orientations of consecutive stimuli (explanatory variable), using the first derivative of a Gaussian (DoG) function (Pascucci et al., 2023). The response variable for the current $n$th trial is formulated as $\mathrm{AE} = D(\hat{\mu}^{(n)}, v^{(n)})$ using (6), where $\hat{\mu}^{(n)}$ is the orientation of the target arm and $v^{(n)}$ is the adjusted orientation in the current trial. To analyze the serial dependence from the target arm in the preceding $(n-1)$th trial, the explanatory variable is formulated as $\Delta\mathrm{T} = D(\hat{\mu}^{(n)}, \hat{\mu}^{(n-1)})$, where $\hat{\mu}^{(n-1)})$ is the orientation of the target arm in the preceding trial. For the non-target arm in the preceding trial, the explanatory variable is formulated as $\Delta\mathrm{NT} = D(\hat{\mu}^{(n)}, \bar{\mu}^{(n-1)})$, where $\bar{\mu}^{(n-1)}$ is the orientation of the non-target arm in the preceding trial. The trial AEs were averaged for each participant. The pairs of the participant-wise AE and Δ were used to fit the DoG function's parameters—amplitude

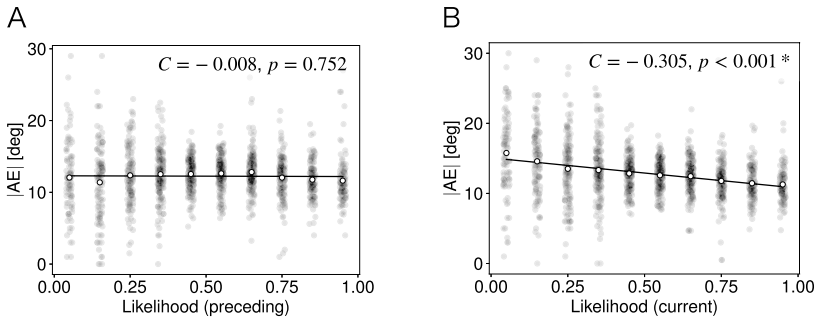



**Fig. 5.** Relationship between action likelihood and AE. **A**. Of the action likelihood in the preceding trial. The black dots show individual participant samples, the white dots show the average across participants, and the solid line represents a linear fit using least squares estimation. To average the values for each participant, the likelihood values for each trial were rounded to the nearest value in the set $\{0.05, 0.15, \ldots, 0.95\}$. **B**. Of the action likelihood in the current trial.

and scale, representing the magnitude and variance of the Gaussian function, respectively—to minimize the prediction error evaluated by the squared error. The SLSQP method, implemented via `scipy.optimize.minimize` in Python, was used for this optimization.

A bootstrap test was employed to evaluate the presence of serial dependence in the DoG fitting results. This test involved random permutation of the pairs of AE and Δ, fitting the DoG model to these permuted pairs, and then calculating the squared error. Serial dependence was inferred if the squared error, $e$, for the original samples was significantly lower than that for the permuted samples. To test for its presence, we repeated the permutation procedure 10,000 times ($M = 10,000$) and computed the squared errors for each run. Consequently, we calculated the $p$-value as $p = K/M$, where $K$ represents the counts of permuted sample sets yielding a squared error lower than $e$.

### 3.2. Results and discussion

Our task (Fig. 4A) required the participants to choose between the black and white arm (the target arm) and adjust for the average orientation of the bars corresponding to the selected arm. The adjustment error (AE) in the orientation, as shown in Fig. 4B, led to the exclusion of samples wherein the AE exceeded the 90th-percentile from the dataset for subsequent analysis. The average AEs before and after the exclusions were $-0.06 \pm 21.66$ and $-0.05 \pm 15.01$, respectively. In our analysis of arm selection, we observed a significant preference for the white arm as the target over the black arm, as shown in Fig. 4C ($t(178) = -5.811$, $p < 0.001$). Moreover, the AE associated with the white arm was lower than that associated with the black arm, as shown in Fig. 4D ($t(178) = 2.899$, $p = 0.004$). For RTs for the orientation adjustment of the two arms, no significant difference was found (Fig. 4E, $t(178) = 0.651$, $p = 0.516$). As the maximum potential rewards for both arms were randomly set, biases in choice and adjustment could not be attributed to the reward structure. Consequently, we inferred the presence of an intrinsic bias toward color in both the selection of the target arm and the accuracy of the orientation adjustment.

We investigated the influence on AE of action likelihood, as estimated by the RL model. The likelihood of an action in the preceding trial did not affect the AE (Fig. 5A). Conversely, the likelihood associated with the action in the current trial significantly affected the error (Fig. 5B). This indicates that the participants made less precise adjustments in orientation during the exploration phase (lower likelihood), compared with the exploitation phase (higher likelihood). This finding contrasts with the results of Experiment 1, which demonstrated no significant effect of action likelihood on AE.

We explored serial dependence across samples filtered by the following conditions:

- Learning phase (Explor/Exploi): According to the RL model fitting, whether a participant was in the exploration (Explor) or exploitation (Exploi) phase during the stimulus presentation in the preceding trial.
- Continuity of the target arm (Keep/Change): Whether the target arm was kept from the preceding trial (Keep) or changed (Change).
- Serial dependence to which orientation (ΔNT/ΔT): Serial dependence, anchored for the orientation of either the non-target arm (ΔNT) or the target (ΔT) in the preceding trial, was assessed.

Fig. 6A shows the correlation between the adjustment error and orientation difference. Our statistical analysis identified serial dependence effects under the eight conditions. The magnitude of these effects, as shown in Fig. 6B, was relatively modest, especially when compared to previous studies such as that conducted by Fischer and Whitney (2014), wherein participants adjusted the orientation of Gabor patches, resulting in a pronounced effect of approximately 8°. The diminished serial effect observed in our study could be attributed to several factors: the nature of the stimulus (randomly oriented short bars versus a single Gabor patch), the presence of distractors (oriented bars representing the non-target arm), and variations in the experimental form (offline face-to-face versus online). The conditions for ΔT indicate an attractive effect, wherein the orientation of the adjustment influences the current adjustment error. These findings align with expectations based on the existing serial dependence literature (Pascucci et al., 2023), suggesting that past percept and/or decisions subtly biased the current orientation adjustment process.

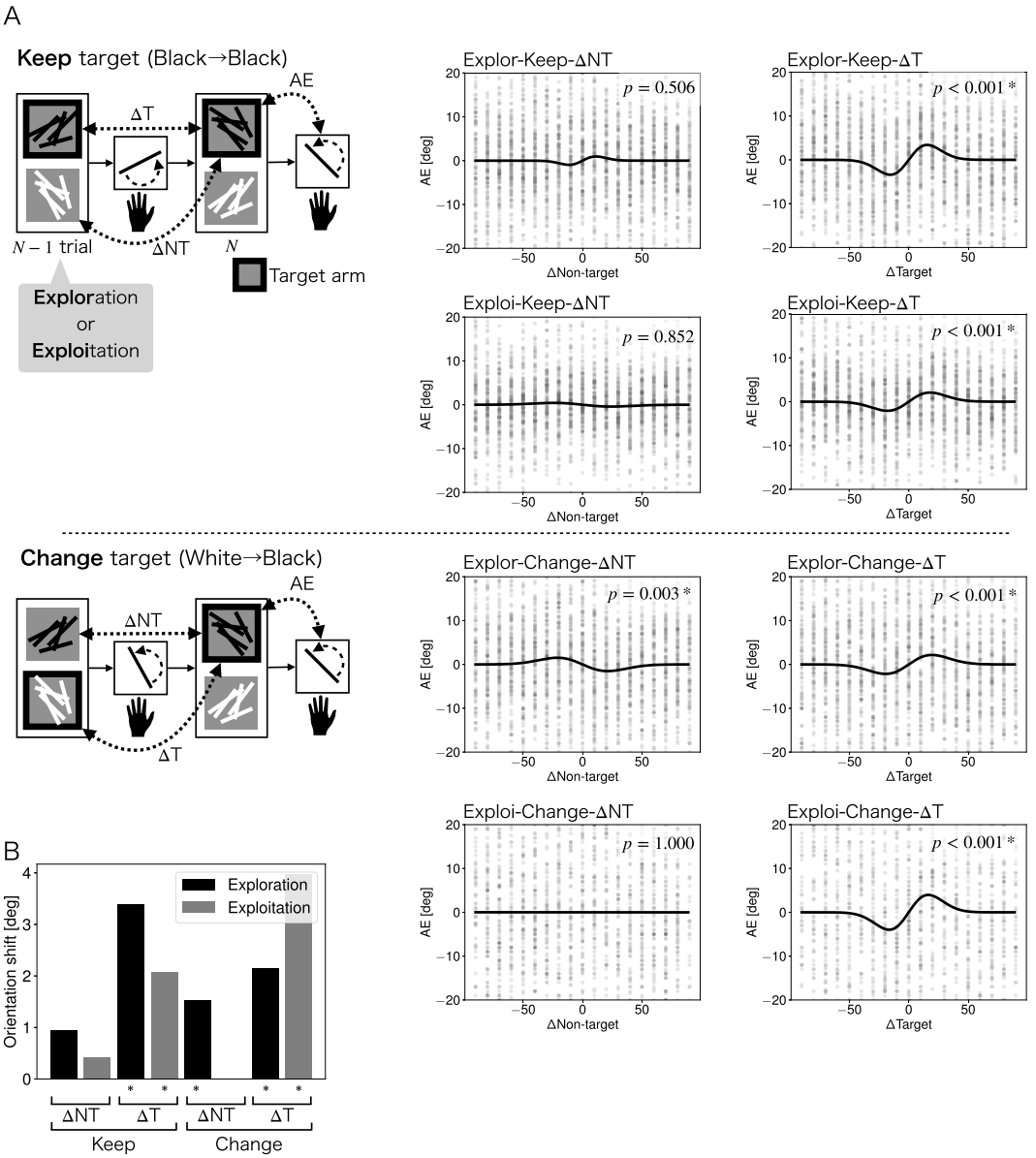



**Fig. 6.** Serial dependence analysis results. **A**. Adjustment errors (AEs) and fitting results. Each dot represents the AE averaged for each participant. The solid lines depict the prediction by the DoG model. The *p*-value in the top-right corner at the panel was adjusted by Bonferroni correction method with eight hypothesises. **B**. Orientation shift deduced from the DoG model's peak.

Serial dependence was also observed under Condition Explor-Change-ΔNT. Conventionally, if the non-target arm does not capture attention, its orientation should not influence the current adjustment. The absence of a serial effect in Conditions Explor-Keep-ΔNT, Exploi-Keep-ΔNT, and Exploi-Change-ΔNT supports the lack of attention to the non-target arm. Nonetheless, the findings for Condition Explor-Change-ΔNT revealed a repulsive effect influenced by the orientation of the non-target arm in the preceding trial. This indicates that participants, particularly during the exploration phase, allocated attention to the non-target arm, even though they did not directly contribute to reward-guided actions.

## 4. General discussion

This study investigated how the learning phase influences visual attention during the observation stage. We utilized the repetition priming effect on the motion direction adjustment of the RDM and the serial dependence effect on the orientation adjustment of the oriented bars to observe visual attention. Our findings revealed that the repetition priming and serial dependence effects were present only during the exploration phase. This indicates that during the exploration phase, the participants paid attention to visual stimuli corresponding to the non-target arm. It appears that the learning phases modulates visual attention deployment: whereas the

exploitation phase focuses attention on target information sources, the exploration phase broadens attention to include surrounding information sources, not just the target.

While there is a rational motivation for focusing on target information sources during the exploitation phase, the motivation for broadening attention during the exploration phase remains an open question. Our results indicate that participants were attentive to the stimulus not corresponding to the target arm in the preceding trial, and this pre-observation drove repetition priming and serial dependence in the subsequent trial (Fischer & Whitney, 2014; Fritsche & de Lange, 2019). However, why did the pre-observation of the non-target arm occur during the exploration phase? Given that participants' primary task was to respond to the property of their chosen target arm, they did not need to gather information on the non-target arm, similar to the exploitation phase. The underlying mechanisms driving this dispersion of attention require further investigation.

The influence of the current trial phase on the adjustment error, observed in Experiment 2, suggests that participants prioritized deploying their attention to the non-target arm. The low adjustment accuracy in the exploration phase (low action likelihood) indicates a voluntary allocation of limited attentional resources (Lavie, 1995; Norman & Bobrow, 1975) to multiple information sources, possibly at the expense of task performance. Conversely, when attentional capacity is sufficient to accurately process the properties of both the target and non-target arms, as observed in Experiment 1, attention distributed to the non-target arm does not compromise adjustment accuracy. Further experiments investigating how the serial effect varies with task difficulty (e.g., changing the stimulus presentation time and coherence of the motion directions of the dots) could provide valuable insights into this hypothesis. Moreover, combining our task with EEG (Cao et al., 2019) and SSVEP frequency tagging (Norcia et al., 2015; Müller et al., 2006; Renton et al., 2021) may enable a deeper investigation into how attention disperses and transitions (Usher & McClelland, 2001) between the target and other arms.

Another difference between Experiments 1 and 2 was the presence of a connection in the stimuli between successive trials. In Experiment 2, the mean orientation of the bars were completely random, with no connection between trials. Conversely, in Experiment 1, the mean motion direction of the RDM was the same as the preceding non-target arm under the Hold trial condition. If the participant noticed this repetition pattern, it might have motivated them to pay attention to the non-target arm. Although the post-test questionnaire indicated that the participants were unaware of the repetition, our experiment cannot rule out the possibility of unconsciousness detection of this regularity in Experiment 1.

Another limitation concerns the procedure for classifying trials into the exploitation and exploration phases. We classified these phases based on action likelihood with a manually adjusted threshold. Our results are dependent on this threshold for trial phase classification. However, the relation between the likelihood and exploratory action is still under discussion (Gershman, 2019; Tomov et al., 2020), and a definitive, reasonable choice has yet to be established. The classification is also depends on RL model used for estimating action likelihood. We implemented a straightforward RL model, assuming that performance on the adjustment task did not influence the learning process related to the bandit problem. However, since the reward is determined by the adjustment error, the confidence level in this adjustment may have affected the update of the value functions (De Martino et al., 2013; Sepulveda et al., 2020; Maldonado Moscoso et al., 2023). To explore this aspect more thoroughly, additional experiments and analyses that integrate confidence into the RL model are required. One promising avenue is adopting a model that incorporates confidence, such as the one proposed by Brus et al. (2021). This would enable a more nuanced understanding of how confidence in perception (i.e., the adjustment task) might interact with learning and decision-making.

In this study, we discovered that the exploration phase led to the dispersion of visual attention. This dispersion, unprompted by rewards in our task, hints at a distinct information-seeking process during the observation stage of the exploration phase, compared with the exploitation phase. Information-seeking behavior during exploration can be characterized as an *active observation*, where the focus is on identifying a target for observation. In other words, decision-making and learning commence with observations. This finding prompts further inquiries into how animals may adapt their information-seeking strategies throughout the learning process, offering a fresh perspective on cognitive behavior and learning mechanisms. Future research in this direction could unveil deeper insights into the dynamic interplay among observation, decision-making, and learning.

## CRediT authorship contribution statement

**Hiroshi Higashi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Data availability

All data and analysis code has been made available in https://github.com/hgshrs/rl_serial.

## Acknowledgements

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI [grant numbers 22H05163 and 24K15047]; and Japan Science and Technology Agency (JST) Advanced International Collaborative Research Program (AdCORP) [grant number JPMJKB2307].

## References


Alais, D., Leung, J., & Van der Burg, E. (2017). Linear summation of repulsive and attractive serial dependencies: Orientation and motion dependencies sum in motion perception. *The Journal of Neuroscience*, *37*(16), 4381–4390. https://doi.org/10.1523/JNEUROSCI.4601-15.2017.

Anderson, B. A. (2016). The attention habit: How reward learning shapes attentional selection. *Annals of the New York Academy of Sciences*, *1369*(1), 24–39. https://doi.org/10.1111/nyas.12957.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–10371. https://doi.org/10.1073/pnas.1104047108.

Anstis, S., & Ramachandran, V. (1987). Visual inertia in apparent motion. *Vision Research*, *27*(5), 755–764. https://doi.org/10.1016/0042-6989(87)90073-3.

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective & Behavioral Neuroscience*, *18*(1), 117–126. https://doi.org/10.3758/s13415-017-0556-2.

Brus, J., Aebersold, H., Grueschow, M., & Polania, R. (2021). Sources of confidence in value-based choice. *Nature Communications*, *12*(7337). https://doi.org/10.1038/s41467-021-27618-5.

Campana, G. (2002). Priming of motion direction and area V5/MT: A test of perceptual memory. *Cerebral Cortex*, *12*(6), 663–669. https://doi.org/10.1093/cercor/12.6.663.

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*, *102*(5), 1076–1087. https://doi.org/10.1016/j.neuron.2019.03.043.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B, Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027.

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.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279.

Dong, D., & Atick, J. (1995). Statistics of natural time-varying images. *Network Computation in Neural Systems*, *6*(3), 345–358. https://doi.org/10.1088/0954-898X/6/3/003.

Easdale, L. C., Le Pelley, M. E., & Beesley, T. (2019). The onset of uncertainty facilitates the learning of new associations by increasing attention to cues. *Quarterly Journal of Experimental Psychology*, *72*(2), 193–208. https://doi.org/10.1080/17470218.2017.1363257.

Failing, M., & Theeuwes, J. (2018). Selection history: How reward modulates selectivity of visual attention. *Psychonomic Bulletin & Review*, *25*(2), 514–538. https://doi.org/10.3758/s13423-017-1380-y.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146.

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077. https://doi.org/10.1038/s41593-019-0518-9.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5), 738–743. https://doi.org/10.1038/nn.3689.

Fritsche, M., & de Lange, F. P. (2019). The role of feature-based attention in visual serial dependence. *Journal of Vision*, *19*(13). https://doi.org/10.1167/19.13.21.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277–286. https://doi.org/10.1037/dec0000101.

Laarni, J. J. (1999). Cues facilitate detection of motion in dynamic random-dot patterns. *Perceptual and Motor Skills*, *88*(1), 129–137. https://doi.org/10.2466/pms.1999.88.1.129.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451–468. https://doi.org/10.1037/0096-1523.21.3.451.

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.

Maldonado Moscoso, P. A., Burr, D. C., & Cicchini, G. M. (2023). Serial dependence improves performance and biases confidence-based decisions. *Journal of Vision*, *23*(7), 5. https://doi.org/10.1167/jov.23.7.5.

Mehlhorn, K., et al. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033.

Müller, M. M., Andersen, S., Trujillo, N. J., Valdés-Sosa, P., Malinowski, P., & Hillyard, S. A. (2006). Feature-selective attention enhances color signals in early visual areas of the human brain. *Proceedings of the National Academy of Sciences*, *103*(38), 14250–14254. https://doi.org/10.1073/pnas.0606668103.

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., & Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, *15*(6), 4. https://doi.org/10.1167/15.6.4.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*(1), 44–64. https://doi.org/10.1016/0010-0285(75)90004-3.

Pascucci, D., et al. (2023). Serial dependence in visual perception: A review. *Journal of Vision*, *23*(1), 9. https://doi.org/10.1167/jov.23.1.9.

Peirce, J., et al. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y.

Pilly, P. K., & Seitz, A. R. (2009). What a difference a parameter makes: A psychophysical comparison of random dot motion algorithms. *Vision Research*, *49*(13), 1599–1612. https://doi.org/10.1016/j.visres.2009.03.019.

Pinkus, A., & Pantle, A. (1997). Probing visual motion signals with a priming paradigm. *Vision Research*, *37*(5), 541–552. https://doi.org/10.1016/S0042-6989(96)00162-9.

Renton, A. I., Painter, D. R., & Mattingley, J. B. (2021). Implicit neurofeedback training of feature-based attention promotes biased sensory processing during integrative decision-making. *The Journal of Neuroscience*, *41*(39), Article 0243. https://doi.org/10.1523/jneurosci.0243-21.2021.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Schacter, D. L., Dobbins, I. G., & Schnyer, D. M. (2004). Specificity of priming: A cognitive neuroscience perspective. *Nature Reviews. Neuroscience*, *5*(11), 853–862. https://doi.org/10.1038/nrn1534.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003.

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *eLife*, *9*. https://doi.org/10.7554/eLife.60705.

Summerfield, C., & Koechlin, E. (2008). A neural representation of prior information during perceptual inference. *Neuron*, *59*(2), 336–347. https://doi.org/10.1016/j.neuron.2008.05.021.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction. Adaptive computation and machine learning*. Cambridge, MA: MIT Press.

Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-15766-z.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550.

van Bergen, R. S., & Jehee, J. F. (2019). Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. *The Journal of Neuroscience*, *39*(41), 8164–8176. https://doi.org/10.1523/JNEUROSCI.3212-18.2019.

Walker, A. R., Luque, D., Le Pelley, M. E., & Beesley, T. (2019). The role of uncertainty in attentional and choice exploration. *Psychonomic Bulletin & Review*, *26*(6), 1911–1916. https://doi.org/10.3758/s13423-019-01653-2.

Walker, A. R., Navarro, D. J., Newell, B. R., & Beesley, T. (2022). Protection from uncertainty in the exploration/exploitation trade-off. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *48*(4), 547–568. https://doi.org/10.1037/xlm0000883.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, *8*(2), 227–233. https://doi.org/10.1016/S0959-4388(98)80144-X.

Xia, Y., Leib, A. Y., & Whitney, D. (2016). Serial dependence in the perception of attractiveness. *Journal of Vision*, *16*(15), 28. https://doi.org/10.1167/16.15.28.