bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Reinforcement meta-learning optimizes visuomotor learning

Taisei Sugiyama[1], Nicolas Schweighofer[2], Jun Izawa[3]*

[1]Empowerment Informatics, University of Tsukuba, Tsukuba, Ibaraki, 305-8573, Japan

[2]Biokinesiology and Physical Therapy, University of Southern California, Los Angeles, CA, 90089-9006 USA

[3]Engineering, Information, and Systems, University of Tsukuba, Ibaraki 305-8573, Japan

*Corresponding author:

Jun Izawa

Faculty of Engineering, Information, and Systems

University of Tsukuba

1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Tel.: +81-29-853-3756; Email: izawa@emp.tsukuba.ac.jp

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Reinforcement learning enables the brain to learn optimal action selection, such as go or not go, by forming state-action and action-outcome associations. Does this mechanism also optimize the brain's willingness to learn, such as learn or not learn? Learning to learn by rewards, i.e., reinforcement meta-learning, is a crucial mechanism for machines to develop flexibility in learning, which is also considered in the brain without empirical examinations. Here, we show that humans learn to learn or not learn to maximize rewards in visuomotor learning tasks. We also show that this regulation of learning is not a motivational bias but is a result of an instrumental, active process, which takes into account the learning-outcome structure. Our results thus demonstrate the existence of reinforcement meta-learning in the human brain. Because motor learning is a process of minimizing sensory errors, our findings uncover an essential mechanism of interaction between reward and error.

Keywords: meta-learning, reinforcement learning, motor learning, cerebellum, basal ganglia, reward, punishment, saving

Learning is considered a skill that can be improved by experience and motivation[1,2]. In fact, animals and humans often exhibit accelerated learning over training sessions[3]. How does the brain learn to learn quickly in variable tasks and environments?

Accelerated learning has often been reported in motor learning tasks where human participants learn to compensate for force or visual perturbations to generate a planned movement trajectory[4-8]. After motor memories formed in the initial learning session have been washed out, learning in the second learning session becomes faster than learning in the first session, a phenomenon that is known as the 'saving effect'[4,9]. In this task, motor learning is considered a

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

process of minimizing sensory prediction errors, i.e., the discrepancy between the generated and predicted movement trajectories is independent of the motivational signals[10-12]. Research has shown that this saving effect is achieved by updating the learning rate (i.e., the policy of how much the motor memory is updated in response to the perceived sensory errors), which is driven by prior experience of the errors[7]. Additionally, recent studies suggest that there is a significant effect of motivational signals on the learning speed[13-17]. How does the brain incorporate motivational signals into the history of errors to regulate learning rates?

One suggested mechanism is a passive, Pavlovian (cue-outcome-based) process in which the valence (reward or punishment) biases the learning rates. For example, in the go and no-go learning task, valence biases the speeds as well as the asymptotes of learning curves[18]. In saccadic adaptation, rewards increase the speeds of adaptation, which is induced by dopaminergic modulation of error signals[19]. While this motivational, hard-wired process might induce biases in error sensitivity of motor learning, it might not explain variations in reward influence of motor learning: some studies show that learning is facilitated by motivational signals[14], while others report no effect[16,17], or even decelerated learning[15]. Thus, these variations imply that another mechanism other than the hard-wired mechanism regulates the speed of learning.

In theory, the ultimate goal of motor learning is to maximize future rewards[20]. Thus, reinforcement learning, which is driven only by reward feedback without any sensory prediction error, also forms motor memory[21-23]. Because the spatial generalization function of learned memory, which reflects properties of neural basis of adaptation, is significantly different for error-based and reward-based learning, two dissociable neural mechanisms are likely involved in error- and reward-based motor learning[24].

However, how these two learning systems interact with each other is still not known. Here, we hypothesized that the integration of error-based and reward-based learning systems reinforces motor learning, which we call 'reinforcement meta-learning.' In contrast to a passive

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

process, this learning process is active and instrumental (action-outcome-based), where a higher-level reinforcement learning mechanism trains a lower-level motor learning mechanism.

This idea of reinforcement meta-learning is based on a theory developed in machine learning where the parameters characterizing learning behaviors such as learning rates are modulated by high-order reinforcement learning[25,26]. This idea can account for changes in learning rates in decision-making tasks in animals and humans[27]. However, the existence of reinforcement meta-learning in the brain has been difficult to verify experimentally because in decision-making tasks, both learning and meta-learning are updated by a reward feedback, obscuring the contribution of each learning layer.

Here, we devise a reinforcement meta-learning task in which the feedback for learning and that for meta-learning are dissociated: sensory error feedback is provided in a motor learning trial, and reward feedback that depends on the rate of motor learning is provided in a subsequent meta-learning trial. If the brain employs reinforcement meta-learning, it integrates sensory error feedback and reward feedback to form associations between these two to regulate how large motor commands should be updated in response to the observed sensory error. Thus, manipulating the relationship between the extent of learning and the reward feedback should influence how motor learning rates are modulated as a result of meta-learning regardless of which reward and punishment are presented. Alternatively, if the modulation of learning rate is hard-wired to valence, such manipulation might not influence learning rates.

Results

Forty one healthy participants gave informed consent before participating in the experiment, which was approved by the Institutional Review Board at the University of Tsukuba. One subject was excluded from the analysis since he reported an explicit strategy to perform task. The participants sat on a chair, held the handle of a robot manipulandum with their right arm, and made shooting-like quick movements by displacing the handle of the robot on a horizontal plane[24].

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A computer projector displayed all visual stimuli onto a flat opaque board that occluded both the manipulandum and the arm. The visuomotor meta-learning paradigm was composed of a sensory-error (S) trial and a monetary feedback (M) trial. In the S trial, the visual cursor was projected on the screen, which provides online feedback of the hand (Figure 1A, left). This cursor was rotated ±7° with respect to the hand movement to induce a sensory prediction error, i.e., the error between the predicted and generated movements. The goal for the shooting was presented with an arc of ±45° instead of a target, and the participants were asked to randomly choose a movement direction and cross somewhere on the arc except the edge, which emphasizes the sensory prediction error and minimizes the extent of how reach error interferes with task performance. This S trial was followed by M trials, in which the small visual target, instead of the arc, was presented, and the participants were asked to shoot at the center of the target as accurately as possible without the cursor feedback. After the shooting movement, the participants received monetary feedback, which was presented as a numerical score (Figure 1A, right). This score was computed as a function of the size of the memory update (learning), which was measured by the aftereffect (i.e., the changes in the reach direction following the previous sensory prediction error). The participants repeated the cycles of one S trial followed by four M trials.

This task structure implements a reinforcement learning problem among trial sequences formalized as a Markov decision process, whereby the agent observes the environmental state, responds according to its policy and is rewarded for the action[28]. In our task, observation of the sensory prediction error (state observation) and the learning rate (policy) yielded a memory update (action), which subsequently generated a monetary feedback (outcome, Figure 1A, bottom). Thus, if the brain learns to maximize the monetary feedback, the action-outcome structure (i.e., how much is gained by how much is learned) should determine the change in learning rate. In addition, if the effect of this action-outcome structure is dominant, the valance which has been considered to influence learning rates[13] does not influence the change in learning rate.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The experiment contained five blocks, each of which comprised a training phase followed by a generalization phase. The training phase had 28 cycles, each containing one S trial and four M trials. The generalization phase included four short sequences of visuomotor rotation trials to assay the generalizability of meta-learned learning rates to a conventional visuomotor learning task (Figure 1B).

An example of the visual rotation pattern (+7° after −7°) is shown in Figure 1C. In the first (baseline) block, no score was given in M trials. Nevertheless, after the sensory prediction error led by visual cursor rotation was observed at the kth trial (S trial), the reach direction was updated to compensate for the given rotation. Thus, the change in movement direction at the kth + 1 trial was significantly different from that at the kth - 1 trial (cyan arrow in Figure 1C, Wilcoxon paired signed-rank tests, $V = 726$, $p < 0.00001$, r=0.70), indicating that the robust aftereffects of memory updates were induced by the sensory prediction errors, which is congruent with previous reports of the roles of sensory prediction errors on updates of motor memory[29].

In subsequent blocks, the participants were randomly assigned to one of four groups (n = 10/group) with different action-outcome structures and valences. In the learn (Lrn) structure, larger aftereffects yielded larger scores, whereas in the not-learn (NLrn) structure, smaller aftereffects yielded larger scores, indicated by the background gradation colors in Figure 1D. The valence determined whether the monetary feedback was positive (reward [Rwd]) or negative (punishment [Pun]). This design required the participants to either learn more to gain more rewards (Lrn-Rwd) and avoid larger punishments (Lrn-Pun) or learn less to gain more rewards (NLrn-Rwd) and avoid larger punishments (NLrn-Pun). To visually observe differences in the regulation of memory updates among these conditions, we focused on the same example rotation pattern (+7° after −7°) as that for the baseline analysis (Figure 1D), and the same Wilcoxon paired signed-rank test was performed. The analysis showed robust aftereffects in the Lrn groups but not in the NLrn groups (Lrn-Rwd: $V = 53$, $p = 0.0059$, r=0.80; Lrn-Pun: $V = 55$, $p = 0.002$, r=0.91; NLrn-Rwd: $V = 37$, $p = 0.38$, r=0.10; NLrn-Pun: $V = 18$, $p = 0.38$, r=0.10). That is, when the

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

memory update led to an increase in rewards and a decrease in punishment, the aftereffect was kept robust; in contrast, when the memory update led to a decrease in rewards and an increase in punishment, the aftereffect was attenuated.

We considered that the participants regulated the learning rate in accordance with the action-outcome structure (Lrn v.s. NLrn). To confirm this, for all combinations of perturbations (+7° after +7°, −7° after −7°, +7° after −7°, and +7° after −7°), we estimated the learning rate, β, by taking the ratio of the memory update (aftereffect) to the sensory prediction error. In other words, we computed how much the reach direction changed relative to the sensory prediction error in an S trial (see Methods for details). The median β was taken for each individual participant and each block for analysis. Figure 2A shows the group mean of individual medians over the blocks for each group and illustrates that the increase or decrease in learning rate over the blocks was different for the groups. To confirm this, we examined the trend of the learning rate over the blocks using a linear mixed-effect model with the action-outcome structure, valence, and block as the fixed effects and participant as the random intercept effect (see Methods for details). This analysis revealed a significant interaction between the action-outcome structure and the block ($t = -2.65$, $p = 0.009$, R^2=0.61). Additionally, we estimated marginal slopes for the change in learning rate over the blocks with respect to the action-outcome structure, which confirmed that the change in learning rate over the blocks was larger in the Lrn groups than in the NLrn groups ($t = -2.61$, $p = 0.0099$). However, we did not find a significant effect of valence ($t = 1.76$, $p = 0.08$). Furthermore, we computed the change in learning rate compared with the baseline for each block and then took the individual participant means of this value across the blocks ($\Delta\beta$, Figure 2B). This result was then analyzed by a reduced linear model with the action-outcome structure and valence as the fixed effects (see Methods for details). Then, we confirmed that the change in learning rate ($\Delta\beta$) was larger in the Lrn group than in the NLrn group ($t = -2.14$, $p = 0.04$, R^2=0.15), while neither valence ($t = 1.21$, $p = 0.23$, R^2=0.15) nor the interaction between action-

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

outcome and valence (t = −0.38, p = 0.70, R^2=0.15) has a significant effect. These results demonstrate that the participants regulated their learning rates according to the action-outcome structure regardless of valence.

If these regulations are achieved by reinforcement learning of learning rates, outcomes that were provided as scores should be maximized. Figure 2C shows the across-subjects mean of the block total score in all conditions. Notably, the baseline was not presented to the participants but was calculated in the same way as the following intervention blocks. As a result of the regulation of the learning rate, participants in all conditions increased their monetary outcome (one-sample Wilcoxon-test; Lrn-Rwd: V = 48, p = 0.04, r=0.56, Lrn-Pun: V = 49, p = 0.03, r=0.61 NLrn-Rwd: V = 55, p = 0.002, r=0.91, NLrn-Pun: V = 49, p = 0.03, r=0.61). These results support the hypothesis that the participants both up- and down-regulated the learning speeds via reinforcement meta-learning: instrumental and active meta-learning processes.

Does this reinforcement meta-learning regulate learning rates of motor learning or something else specific for this training task? Subsequently, after the training phase, the participants experienced the generalization phase, which involved a conventional visuomotor rotation task in which both the visual target and the hand cursor were presented and no monetary feedback was provided (Figure 1B). Then, we investigated the learning rate in the generalization phase to determine whether the effect of the action-outcome structure in the reinforcement meta-learning was transferred to a conventional visuomotor rotation task. Additionally, we investigated time-dependent changes within the generalization phase to determine the temporal robustness of the effect. To do so, we estimated the learning rate, β, for each individual, block, and sequence, and then the individual mean change from baseline across the blocks, $\Delta\beta$, was calculated for each sequence. Figure 3A shows the group means for individual mean change over the sequences for the Lrn and NLrn groups. The individual mean change in learning rate was analyzed with a linear mixed-effect model with the action-outcome structure, the sequence, their interaction as

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fixed effects and the participant as the random intercept effect (see Methods for details). The analysis revealed a significant effect of the action-outcome structure ($t = -2.08$, $p = 0.04$, R^2=0.67) and its interaction with the sequence ($t = 2.58$, $p = 0.01$, R^2=0.67). Furthermore, the individual mean changes in learning rate for the training phase correlated with those in the first half (Figure 3B, $R = 0.43$, $t_{38} = 2.90$, $p = 0.006$) but not with those in the second half (R=0.25, $t_{38} = 1.61$, $p = 0.12$) of the generalization phase. Notably, although the change in learning rate was greater in the Lrn groups than in the NLrn groups in the first half (vertical axis of Figure 3B, $t = -2.78$, $p = 0.008$, $d = .88$), this phenomenon was not observed for the second half ($t = -0.74$, $p = 0.46$, $d = .23$). These results demonstrate that the regulation of the learning rate by reinforcement meta-learning in the training phase was generalized to the learning rate in motor learning for conventional visuomotor rotation tasks. This suggests that what is updated in the training phase overlaps with the learning rate of visuomotor learning.

Discussion

We found that by presenting rewards as a function of the amount of memory update that was induced by sensory prediction errors, the brain both up- and down-regulated the learning rates to increase rewards as well as to avoid punishment. This effect was gradually developed over the training sessions, and the directionality of regulation (i.e., learn or not-learn) was determined by the action-outcome structure, not by valence. These results suggest that reinforcement learning is employed to regulate learning rates. Furthermore, by examining the generalization of these regulations from the training task to the conventional visuomotor rotation task, we found that these meta-learning effects were transferred to motor learning. This observation indicates that there is an overlapping neural basis between reinforcement meta-learning and visuomotor learning. Thus, our results demonstrate the existence of a reinforcement meta-learning mechanism for motor learning in the human brain.

In machine learning studies, automatic tuning of learning parameters has been a long-

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

standing problem[30]. A computational model of biological reinforcement learning suggests that, in the brain, neuromodulators such as noradrenaline, acetylcholine, and noradrenaline adjust learning parameters of reinforcement learning[31] via reward-based modulation of these parameters[32]. Recent algorithms of meta-learning in machine learning studies have highlighted a hierarchical structure composed of two reinforcement learning systems: while a low-order reinforcement learning optimizes the weight parameters for action selections in a single learning episode, a high-order reinforcement learning optimizes meta-parameters of the low-order reinforcement learning network to maximize rewards across multiple learning episodes[25,26]. Although the neural implementation of such reinforcement meta-learning was recently discussed[27], there was no experimental examination. Here, we devised a meta-learning paradigm in which reward feedback for meta-learning, which was provided independently from the sensory error feedback for motor learning, was manipulated as a function of the learning rate of motor learning. Our data demonstrated that while motor memory was updated to minimize the given sensory prediction errors in a single trial, the learning parameter of motor learning was updated to maximize rewards over multiple trials.

The observed meta-learning effect may account for previous reports of learning-rate flexibility during motor learning tasks[6,7,13,33]. For instance, the learning rate increases over the sessions when the perturbation is relatively constant but decreases when the direction of perturbation frequently changes. Importantly, a typical learning paradigm with constant perturbation shares the same action-outcome structure as that for Lrn-Pun, whereby a memory update attenuates movement errors on the next trial, which can be considered avoidance of aversive outcomes[34,35]. In this case, reinforcement meta-learning accelerates learning rates to quickly reduce errors in re-learning (e.g., saving)[6] or learning with additional punishment signals[13]. Conversely, a non-typical learning paradigm with rapidly changing perturbation shares the same action-outcome structure as NLrn-Pun because the memory update increases errors in the next trial. In this case, reinforcement meta-learning decreases the learning rate. Thus,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reinforcement meta-learning explains how the statistics of the perturbation affect learning rates[7,33].

According to a conventional theory of motor learning, the brain updates motor commands independent of reward to minimize sensory prediction errors[36]. Although motor memory has also recently been found to be updated by rewards, the neural basis of this reward-based motor learning is likely distinct from that of sensory error-based motor learning[24,37]. According to these previous studies, rewards might not interact with sensory prediction errors during motor learning. Here, our data demonstrate that the sensory prediction errors and the rewards presented in separated trials were integrated at the higher-level motor learning system, i.e., meta-learning, to regulate learning rates.

In theory, to establish reinforcement meta-learning for motor learning, the sensitivity to sensory prediction errors should be evaluated by the rewards. Thus, our results suggest a close interaction between rewards and sensory prediction errors during motor learning. Because research evidence suggests the involvement of cortico-basal ganglia networks in reinforcement meta-learning[27] and the cerebellum in sensory error-based learning[38,39], reinforcement meta-learning for motor learning is likely mediated by the functional connectivity between these two learning systems. Thus, the anatomical projections between the basal ganglia and the cerebellum could have a computational role in reinforcement meta-learning[40]. This possibility is further supported by recently reported reward-related signals in cerebellar inputs[41] and outputs[42] during motor control tasks. We suggest that these interactions between the basal ganglia and the cerebellum play a key role in optimizing learning parameters for motor learning via reinforcement learning.

Acknowledgments

This work was supported by KAKENHI (grant numbers 17H06023 and 19H04977). T.S. is grateful for financial support from the JSPS Research Fellowship for Young Scientist and KAKENHI (19J20366).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Author Contributions

Conceptualization, T.S. and J.I.; methodology, T.S. and J.I.; investigation, T.S. and J.I.; formal analysis, T.S. and N.S.; writing – original draft, T.S., and J.I.; writing – review & editing, T.S., N.S., and J.I.; funding acquisition, J.I.

Declaration of Interests

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

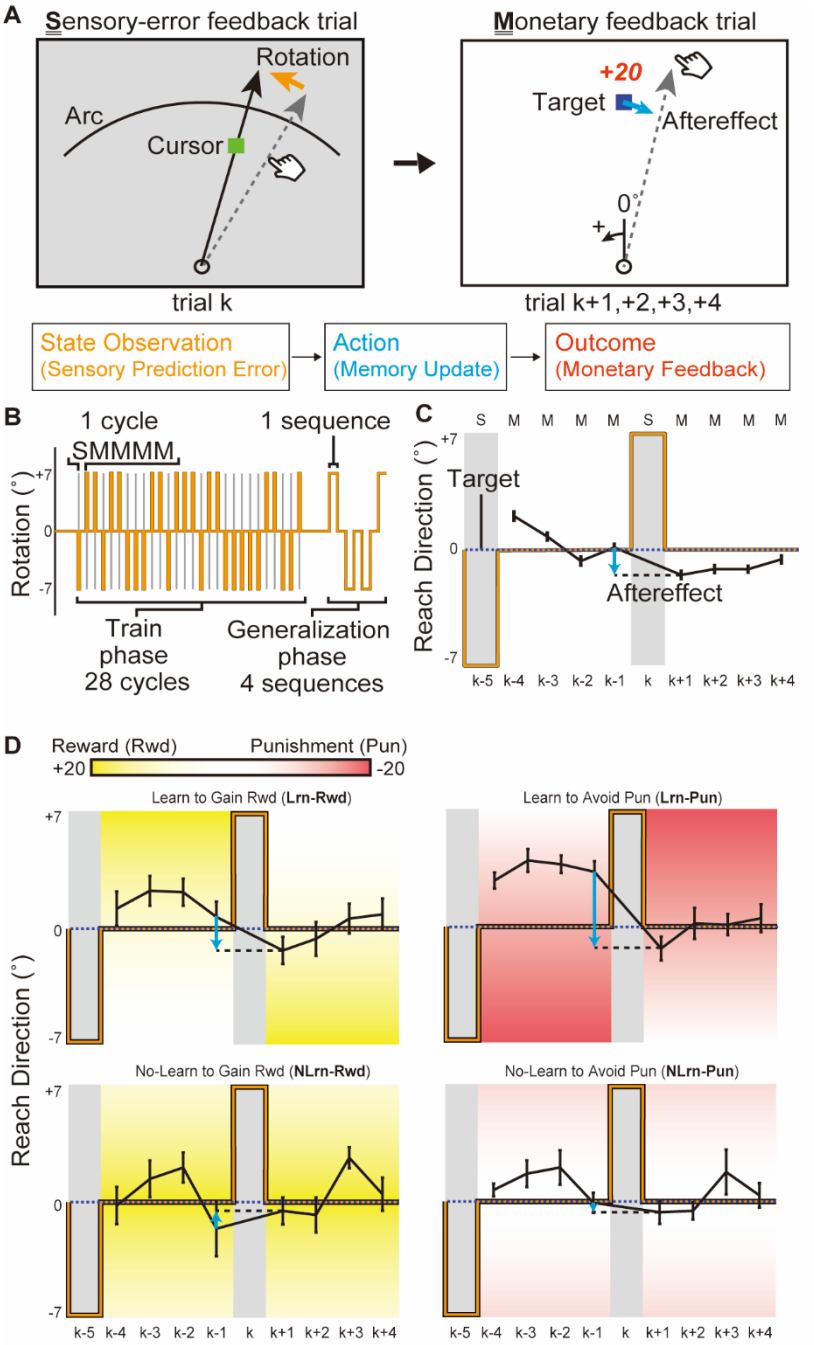

Figure 1. Reward-based visuomotor meta-learning paradigm. (A) Training paradigm. Cursor rotation (orange arrow)-induced sensory prediction error observation (state) in the sensory error (S) trial. The aftereffect size (action, cyan arrow) was scored (outcome) in the monetary feedback (M) trial. (B) One block consisted of the training phase, which included 28 cycles of 1 S trial

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

followed by 4 M trials. After brief washout (null) trials, the generalization phase was composed of four short sequences of visual rotation trials without scores. (C) Plot of reach direction w.r.t. the target direction (memory) for one example rotation pattern in baseline. The aftereffect (cyan arrow) was developed in the opposite direction to the rotation (orange line) in the S (k^{th}) trial, showing a motor memory update. (D) Memory profile for the same rotation pattern for each group. The scores improved with memory updates toward the yellow zone (large reward) or away from the red zone (large punishment). The error bars indicate standard error of the mean (SEM).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

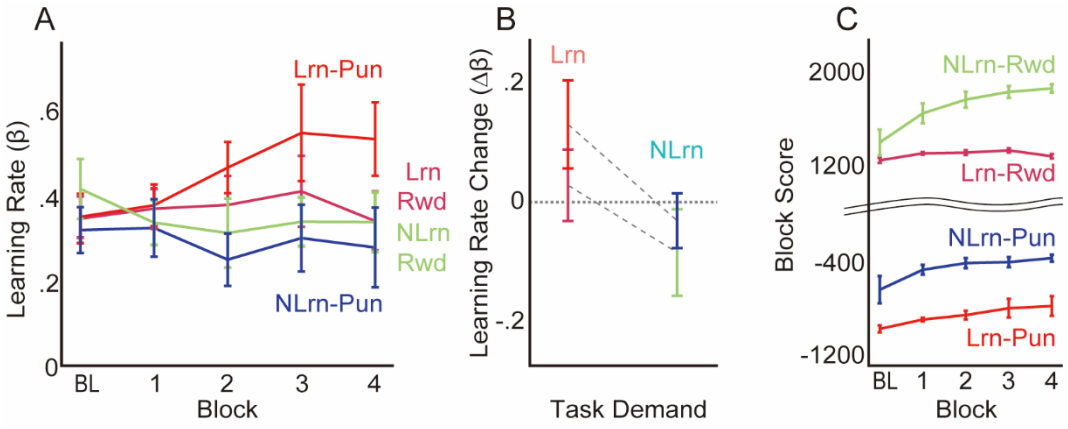

Figure 2. Mean performance scores and learning rates. (A) Estimated changes in learning rates by block. (B) The learning rates changed more in the Lrn group than in the NLrn group. The error bars indicate SEM. (C) Total scores for each block. The baseline (BL) scores were calculated and are shown here.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

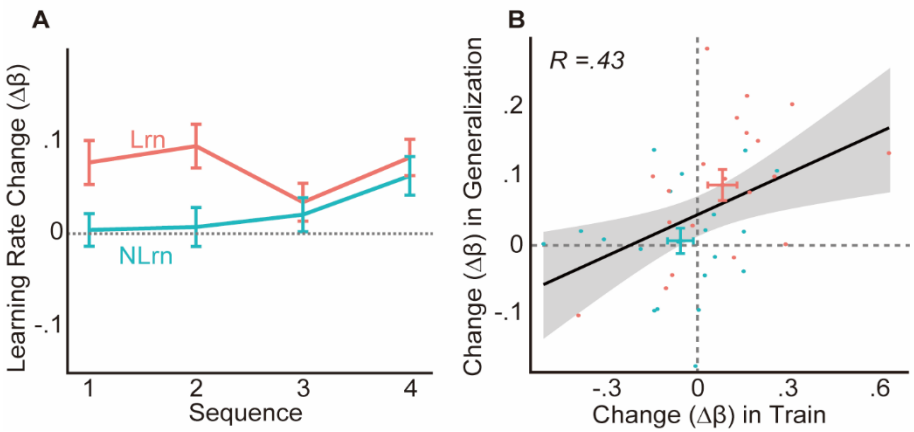

Figure 3. Mean learning-rate changes in the generalization phase. (A) Mean learning-rate changes across the blocks for each sequence. (B) The learning-rate changes in the training phase corresponded with those in the first half of the generalization phase. Each dot represents the data from an individual participant. The error bars indicate SEM.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Participants

Forty one right-handed participants (25 males; aged 19–37 years, μ = 24) volunteered for this experiment. Their handedness was confirmed by the Edinburgh Handedness Inventory, and they reported no history of neurological or motor disorders. We excluded one subject who reported an explicit strategy to perform the task from the analysis. They were paid 1,640 JPY for their participation, with an additional performance-based compensation up to 1,000 JPY.

Task design

General

Participants performed the task using a robot manipulandum[24] that moved only in the horizontal plane. They sat on a chair and held the robot handle in their right hand. A horizontal, flat, opaque board covered the task space, occluding the hand and forearm. A computer projector was fixed above and displayed visual information on the board.

In each trial, participants made a rapid shooting movement when a target appeared 10 cm away from the starting point. To control for use-dependent learning[43], the target location was pseudo-randomly selected from 1 of 7 locations: -15°, -10°, -5°, 0° (directly in front of the participant), 5°, 10°, and 15°. The counter-clockwise direction was defined as positive in the angular coordinates (**Figure 1A**). To maintain similar kinematics across trials, “Too Fast” or “Too Slow” was displayed as a warning when the movement duration was <200 ms or >300 ms, respectively.

Trial type and task schedule

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

There were four trial types: Null, Sensory-error feedback (S), Monetary feedback (M), and Generalization. In Null trials, participants made shooting movements toward the targets with veridical online cursor feedback. In S trials, the targets were replaced with an arc spanning ±45°, centered on the home position and with 10 cm radius (**Figure 1A**) [44]. Participants were asked to cross the arc while trying to distribute their reach direction across S trials. Online cursor feedback was rotated ±7° from hand movements to induce errors between the predicted hand position and the visual feedback. The rotation direction was pseudo-randomly selected (**Figure 1B**). Note that the use of an arc instead of a target minimizes the extent of how reach errors interfere with task performance and, thus, emphasizes errors between the observed and predicted hand movements, i.e., sensory predictor errors. In M trials, participants made shooting movements toward targets without online cursor feedback. Upon movement completion, monetary feedback was presented as a numerical score above the target (**Figure 1A**). Scores were computed based on the reach direction from the target, with different computations across the experimental conditions (see below). Finally, Generalization trials were identical to Null trials except that, similar to S trials, online cursor feedback was rotated ±7° from the hand movement, as in conventional visuomotor rotation tasks[45].

Participants performed first one Baseline block, in which no monetary feedback was provided, and then four blocks with monetary feedback. Each block comprised a Train phase with 28 cycles of 1 S trial and 4 M trials, followed by a Generalization phase. Brief washout blocks of 14 Null trials were inserted before each phase. The Generalization phase comprised 4 Sequences of 5 consecutive cursor rotation trials, with 5 Null trials between each Sequence (**Figure 1B**). The aim of the Generalization phase was to test whether the participants generalized the meta-learning effect formed in the Train phase to a conventional visuomotor learning task.

Estimation of learning rates in the Train phase

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We estimated the learning rates in the Train phase with a simple first order state-space model of memory updates to track transitions in reach direction in response to the experienced cursor rotation, as in previous work[7]. In this framework, motor learning is considered a process that estimates perturbations imposed in the task environment. Specifically in the visuomotor rotation task, the executed motor plan $u^{(t)}$ at trial *t* determined the direction of the hand movement $h^{(t)}$,

$$h^{(t)} = u^{(t)}. \qquad (1)$$

While the hand movement was not directly observable for participants, a cursor projected on the screen provided online feedback of the hand motion while the visual rotation $p^{(t)}$ was imposed between the hand movement and the visual cursor:

$$c^{(t)} = h^{(t)} + p^{(t)}. \quad (2)$$

The brain may predict hand movement direction $\hat{h}^{(t)}$ based on the estimation of the perturbation $x^{(t)}$ and the efference copy of the motor plan $u^{(t)}$:

$$\hat{h}^{(t)} = u^{(t)} + x^{(t)}. \quad (3)$$

To minimize the sensory prediction error $c^{(t)} - \hat{h}^{(t)}$, the brain updates the estimate of the perturbation with the following learning rule:

$$x^{(t+1)} = x^{(t)} + \beta^{(t)}(c^{(t)} - \hat{h}^{(t)}), \qquad (4)$$

where the learning rate β characterizes the rate of learning.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here, we estimated the learning rate β in the Train phase using the data of hand directions in a triplet of M, S, and M trials, following previously developped methods[46]. That is, considering a certain trial of S trials at the trial $t = k$, $\beta^{(k)}$ was estimated using the measured hand movement direction $h^{(k)}$ and the given cursor rotation $p^{(k)}$ in S trials at $t = k$, $h^{(k-1)}$ and the presented target direction $T^{(k-1)}$ in M trials at $t = k - 1$, as well as $h^{(k+1)}$ and $T^{(k+1)}$ in M trials at $t = k + 1$. Because in M trials, the visual target T was presented and the participants were explicitly instructed to cross the target with their unseen but estimated hand position, the participants' estimation of their hand direction should closely match the target direction. We thus assumed $T^{(k-1)} = \hat{h}^{(k-1)}$ and $T^{(k+1)} = \hat{h}^{(k+1)}$. From (1) and (3), we have $x^{(k-1)} = T^{(k-1)} - h^{(k-1)}$ and $x^{(k+1)} = T^{(k+1)} - h^{(k+1)}$. Importantly, because the cursor was not given to participants in M trials at $t = k - 1$, the sensory prediction error $c^{(k-1)} - \hat{h}^{(k-1)}$ was absent and thus, no memory update was engaged. Hence, we assumed $x^{(k)} = x^{(k-1)}$. For subsequent S trials at $t = k$, because the sensory prediction error $c^{(k)} - \hat{h}^{(k)}$ was present, the memory was updated in accordance with (4).

Using (4), we estimate $\beta^{(k)}$ by:

$$\beta^{(k)} = \frac{x^{(k+1)} - x^{(k)}}{c^{(k)} - \hat{h}^{(k)}}. \qquad (5)$$

According to (1), (2) ,(3), and (5), we have:

$$\beta^{(k)} = \frac{x^{(k+1)} - x^{(k)}}{(u^{(k)} + p^{(k)}) - (u^{(k)} + x^{(k)})}.$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Then, by applying $x^{(k)} = x^{(k-1)}$, we have:

$$\beta^{(k)} = \frac{x^{(k+1)} - x^{(k-1)}}{p^{(k)} - x^{(k-1)}}.$$

Finally, by substituting $x^{(k-1)}$ by $T^{(k-1)} - h^{(k-1)}$ and $x^{(k+1)}$ by $T^{(k+1)} - h^{(k+1)}$, we have

$$\beta^{(k)} = \frac{(T^{(k+1)} - h^{(k+1)}) - (T^{(k-1)} - h^{(k-1)})}{p^{(k)} - (T^{(k-1)} - h^{(k-1)})}. \qquad (6)$$

Because we measured $T^{(k+1)} - h^{(k+1)}$ and $T^{(k-1)} - h^{(k-1)}$ in M trials and manipulated $p^{(k)}$ as a cursor rotation in S trials, $\beta^{(k)}$ can be estimated for each triplet of M-S-M trials in the Train phase. Note that the denominator $p^{(k)} - (T^{(k)} - h^{(k+1)})$ occasionally approaches 0, resulting in an unreliable estimation of β. To prevent this, we used the median instead of the mean for each Train phase for each participant before performing statistical analyses.

A potential limitation of the above method is that it does not account for the effect of another type of memory that is updated by reward signals[24], which might be formed in M trials at $t = k - 1$. If this reward-based motor memory exists, it affects the measured reach direction in subsequent S trials at $t = k$ and M trials at $t = k + 1$. This could potentially bias the approximation of the sensory prediction error $p^{(k)} - x^{(k-1)}$ by $p^{(k)} - (T^{(k-1)} - h^{(k-1)})$ and thus, bias the estimation of β. However, according to the theory of reward-based motor memory[24], the directions of the bias is opposite in the two cases of perturbation sequences: whether the direction of the perturbation switches or remains the same. Because the perturbation sequence was pseudorandomized, this bias should be predominantly eliminated when we calculate the median over each block. If a small effect of this bias remains, it would likely lead to an underestimation of β for Lrn groups and

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

overestimation for NLrn groups; however, this was not observed in our data. As a final analysis of this potential confounding factor, we examined the generalization of the meta-learning in the Train phase to a conventional visuomotor rotation task where no monetary feedback was given.

Estimation of learning rates in the Generalization phase

Using the state-space model of the memory update (1)-(4), the learning rate was estimated via least square to fit each subject's data of measured hand directions. In the Generalization phase, we assumed that the participants aimed at the presented target $T^{(t)}$ with their prediction of the hand direction $\hat{h}^{(t)}$, so that $\hat{h}^{(t)}$ closely matched with $T^{(t)}$ (i.e, $\hat{h}^{(t)} = T^{(t)}$). They observed the error between the cursor and the prediction of the hand $c^{(t)}$- $\hat{h}^{(t)}$. Thus, the state-space model of the memory update (4) provides us the simulated sequences of $x^{(t)}$ over trials using the measured sequence of $c^{(t)} - T^{(t)}$ for a given β. The model also generates simulated sequence of participants' reach error $E_S{}^{(t)} = T^{(t)} - h^{(t)}$ in accordance with (1) and (3). In the experiment, we measured the actual participants' reach error $E_M{}^{(t)} = T^{(t)} - h^{(t)}$. Thus, following a previous method[4], we estimated the learning rate β that minimized the sum of the least square error between $E_S{}^{(t)}$ and $E_M{}^{(t)}$ over each step-perturbation sequence of the three Null trials and the five cursor rotation trials.

Experiment groups and score calculations

There were two independent variables, Action-Outcome structure and Valence, respectively, with two levels for each. This resulted in four experiment groups (**Figure 1D**). The Action-Outcome structure determined how the monetary feedback was computed in four consecutive M trials ($t = k+i, i = \{1,2,3,4\}$) after the S trial ($t = k$) both as a function of the memory x^{k+i}, measured as

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a reaching angle with respect to the target $T^{(k+i)} - h^{(k+i)}$, and as a function of the given perturbation $p^{(k)}$. This regulated whether the memory updates in response to the sensory prediction error $\beta^{(k)}(c^{(k)} - \hat{h}^{(k)})$ were encouraged (Lrn) or discouraged (NLrn). Specifically, for Lrn, larger aftereffects of the exposure to the cursor rotation corresponded to larger scores, whereas, for NLrn, smaller aftereffects corresponded to larger scores. In contrast, Valence determined whether monetary feedback was positive (i.e., reward) or negative (i.e., punishment). Therefore, participants could learn more from sensory prediction error to gain more rewards (Lrn-Rwd) or to avoid larger punishments (Lrn-Pun), allowing them to improve their scores with greater memory updates. Alternatively, they could learn less from sensory prediction error to gain more rewards (NLrn-Rwd) or to avoid larger punishments (Lrn-Pun), allowing them to improve their score with less memory updates.

The score ranges for the Rwd and Pun groups were set to [0,20] and [-20,0], respectively. The Lrn-Rwd/Pun group earned the highest score by showing 100% or more memory for the last experienced rotation in the previous S trial (i.e., $\frac{T^{(k+1)}-h^{(k+1)}}{p^{(k)}} \geq 1$ in i^{th} M trial within the same cycle), and the score was reduced by 1 point for every 10% less memory until the lowest score (i.e., -100% learning) was reached. The Lrn-Rwd/Pun group earned the highest score by showing 0% memory, (i.e. $T^{(k+i)} - h^{(k+i)} = 0$), and the score was reduced by 1 point for every 10% more/less memory until the lowest score was reached. These were represented as background color patterns in **Figure 1D**.

Instructions

Before the task, participants were instructed about the experimental flow for the Train phase and the stimuli and feedback in the S and M trials. They were also explicitly informed that the total score would determine their additional monetary compensation and that the task goal was to

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

maximize their compensation by crossing the target with their hand as closely as possible. In the Rwd conditions, they were told that additional compensation was initially minimum (0 JPY) and accumulated throughout the task. In the Pun condition, participants were told that additional compensation was initially maximum (1,000 JPY) and subtracted throughout the task. In addition, in the S trial, they were told to vary movement directions from trial to trial. This was to avoid the formation of use-dependent behavior[43]. In addition, they were not informed of visual rotation or the relationship between their reach direction and score size to prevent the use of cognitive strategy[47]. Their unawareness of visual rotation throughout the task was confirmed by a written questionnaire after the task was completed asking if they felt a discrepancy between the hand and the cursor during movements.

Statistical analyses

Removal of target location-dependent bias

Due to a gap in height between the physical hand position and the cursor projected on the screen, a bias between the target direction and the reach direction was inevitable. To remove this bias from the analysis, we calculated the mean reach direction for each target location in the Baseline and subtracted from it in the Train phase for all blocks.

Memory update following a specific visual rotation

In **Figure 1C,** individual mean reach directions with respect to target direction, $T-h$, in the k-1th and k+1th trials were compared by Wilcoxon paired signed-rank test to examine the aftereffect (i.e., if the reach direction changed after the participants experienced a sensory prediction error in the kth trial). Because there was no difference across groups in Baseline, all individual data were combined (n = 40). The same test was performed for memory updates in the other four blocks on

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

individual mean differences of $T-h$ in the k-1th and k+1th trials across the blocks for each group (**Figure 1D**, n = 10/group).

Learning rate in the Train phase

We first used a linear mixed effect model (LMEM) with Action-Outcome structure, Valence, and Block as the fixed effects and Participant as the random intercept effect to examine the change in β over the course of the task, as shown in **Figure 2A**. All interactions between the fixed effects were included. In R with lme4 package[48], the model was written as: $lmer(\beta \sim Action\text{-}Outcome * Valence * Block + (1|Participant))$. In addition, to estimate the marginal slopes of the learning rate change over blocks with respect to Action-Outcome or Valence, we then developed a reduced model without interactions between Action-Outcome and Valence. The model was written as: $lmer(\beta \sim (Action\text{-}Outcome + Valence) * Block + (1|Participant))$.

Then, to examine the overall change in learning rate across the task, the change in β from Baseline ($\Delta\beta$) was calculated for each block and then averaged across blocks for each participant and fit with the following model: $lm(\Delta\beta \sim Action\text{-}Outcome * Valence)$. Note that because one data point was obtained for each participant, no random effects were included in this model.

Block total score performance

To evaluate whether score performance improved, we calculated the total score for each block including Baseline where the score was calculated in the same manner without being presented to the participants, as shown in **Figure 2C**. Then, the change from Baseline was calculated and averaged across blocks for each participant. The change for each group was analyzed by one-

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sample Wilcoxon signed-rank test with a hypothetical mean of 0 (n = 10/group).

Learning rate in the Generalization phase

The change in learning rate $\Delta\beta$ in the Generalization phase was analyzed similarly to that in the Train phase. Again, individual mean change from Baseline across blocks was calculated by subtracting the Baseline average β across Sequence from β in the subsequent blocks. Then, $\Delta\beta$ was analyzed with a LMEM, with Action-Outcome structure, Sequence, and their interaction, as fixed effects and Participant as random effect intercepts: $lmer(\Delta\beta \sim Action\text{-}Outcome * Sequence + (1|Participant))$. Note that Valence was removed from the model based on the analysis of the change in β in the Train phase, which failed to show its significant effect.

In addition, to examine time-dependent changes within the Generalization phase more closely, the data were separated into the first and second half of the phase. To examine if the change in each half of the Generalization phase reflected the change in the Train phase, a Pearson correlation analysis was then performed on the changes between the Train and the first/second half of Generalization phases. Finally, $\Delta\beta$ from each half were analyzed with a linear model with Action-Outcome as the only factor.

Statistical package, coding, and validation of linearity

All data processing and statistical analyses were performed in R version 3.5.1 using the lm, lme4[48], lmerTest[49], and emmeans[50] packages. For the linear model analyses, effect coding was used to represent the categorical variables (i.e., Action-Outcome and Valence), except in estimation of marginal slopes where dummy coding was used, and the numerical variables (i.e., Block and Sequence) were represented in z-score (i.e., centered and normalized). Lrn and Rwd were set as reference in coding (i.e., negative value in effect coding and zero in dummy coding). The

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assumption of linearity was validated by (1) confirming by log-likelihood ratio test on BIC that a model that treats numerical variables as categorical does not produce significantly better fit than a model that treats them as continuous, and (2) confirming the normality of residuals by Shapiro-Wilk test.

References

1 Hattie, J., Biggs, J. & Purdie, N. Effects of learning skills interventions on student learning: A meta-analysis. *Rev Educ Res* 66, 99-136, doi:Doi 10.3102/00346543066002099 (1996).

2 Brown, A. L. & Kane, M. J. Preschool children can learn to transfer: learning to learn and learning from example. *Cogn Psychol* 20, 493-523 (1988).

3 Harlow, H. F. Physiological psychology. *Annu Rev Physiol* 11, 269-296, doi:10.1146/annurev.ph.11.030149.001413 (1949).

4 Smith, M. A., Ghazizadeh, A. & Shadmehr, R. Interacting adaptive processes with different timescales underlie short-term motor learning. *PLoS Biol* 4, e179, doi:10.1371/journal.pbio.0040179 (2006).

5 Leow, L. A., Loftus, A. M. & Hammond, G. R. Impaired savings despite intact initial learning of motor adaptation in Parkinson's disease. *Exp Brain Res* 218, 295-304, doi:10.1007/s00221-012-3060-5 (2012).

6 Kojima, Y., Iwamoto, Y. & Yoshida, K. Memory of learning facilitates saccadic adaptation in the monkey. *J Neurosci* 24, 7531-7539, doi:10.1523/JNEUROSCI.1741-04.2004 (2004).

7 Herzfeld, D. J., Vaswani, P. A., Marko, M. K. & Shadmehr, R. A memory of errors in sensorimotor learning. *Science* 345, 1349-1353, doi:10.1126/science.1253138 (2014).

8 Krakauer, J. W., Ghilardi, M. F. & Ghez, C. Independent learning of internal models for kinematic and dynamic control of reaching. *Nat Neurosci* 2, 1026-1031, doi:10.1038/14826 (1999).

9 Medina, J. F., Garcia, K. S. & Mauk, M. D. A mechanism for savings in the cerebellum. *J Neurosci* 21, 4081-4089 (2001).

10 Tseng, Y. W., Diedrichsen, J., Krakauer, J. W., Shadmehr, R. & Bastian, A. J. Sensory prediction errors drive cerebellum-dependent adaptation of reaching. *J Neurophysiol* 98, 54-62, doi:10.1152/jn.00266.2007 (2007).

11 Wolpert, D. M., Doya, K. & Kawato, M. A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biol Sci* 358, 593-602, doi:10.1098/rstb.2002.1238 (2003).

12 Cullen, K. E. & Brooks, J. X. Neural correlates of sensory prediction errors in monkeys: evidence for internal models of voluntary self-motion in the cerebellum. *Cerebellum* 14, 31-34, doi:10.1007/s12311-014-0608-x (2015).

13 Galea, J. M., Mallia, E., Rothwell, J. & Diedrichsen, J. The dissociable effects of punishment and reward on motor learning. *Nat Neurosci* 18, 597-602, doi:10.1038/nn.3956 (2015).

14 Nikooyan, A. A. & Ahmed, A. A. Reward feedback accelerates motor learning. *J Neurophysiol* 113, 633-646, doi:10.1152/jn.00032.2014 (2015).

15 van der Kooij, K. & Overvliet, K. E. Rewarding imperfect motor performance reduces adaptive changes. *Exp Brain Res* 234, 1441-1450, doi:10.1007/s00221-015-4540-1

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2016).

16 Song, Y. & Smiley-Oyen, A. L. Probability differently modulating the effects of reward and punishment on visuomotor adaptation. *Exp Brain Res* 235, 3605-3618, doi:10.1007/s00221-017-5082-5 (2017).

17 Quattrocchi, G., Greenwood, R., Rothwell, J. C., Galea, J. M. & Bestmann, S. Reward and punishment enhance motor adaptation in stroke. *J Neurol Neurosurg Psychiatry* 88, 730-736, doi:10.1136/jnnp-2016-314728 (2017).

18 Guitart-Masip, M. *et al.* Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage* 62, 154-166, doi:10.1016/j.neuroimage.2012.04.024 (2012).

19 Kojima, Y. & Soetedjo, R. Selective reward affects the rate of saccade adaptation. *Neuroscience* 355, 113-125, doi:10.1016/j.neuroscience.2017.04.048 (2017).

20 Izawa, J., Rane, T., Donchin, O. & Shadmehr, R. Motor adaptation as a process of reoptimization. *J Neurosci* 28, 2883-2891, doi:10.1523/JNEUROSCI.5359-07.2008 (2008).

21 Therrien, A. S., Wolpert, D. M. & Bastian, A. J. Effective reinforcement learning following cerebellar damage requires a balance between exploration and motor noise. *Brain* 139, 101-114, doi:10.1093/brain/awv329 (2016).

22 Pekny, S. E., Izawa, J. & Shadmehr, R. Reward-dependent modulation of movement variability. *J Neurosci* 35, 4015-4024, doi:10.1523/JNEUROSCI.3244-14.2015 (2015).

23 Manley, H., Dayan, P. & Diedrichsen, J. When money is not enough: awareness, success, and variability in motor learning. *PLoS One* 9, e86580, doi:10.1371/journal.pone.0086580 (2014).

24 Izawa, J. & Shadmehr, R. Learning from sensory and reward prediction errors during motor adaptation. *PLoS Comput Biol* 7, e1002012, doi:10.1371/journal.pcbi.1002012 (2011).

25 Finn, C., Xu, K. & Levine, S. Probabilistic Model-Agnostic Meta-Learning. *Adv Neur In* 31 (2018).

26 Duan, Y. *et al.* Rl $^ 2$: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779* (2016).

27 Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci* 21, 860-868, doi:10.1038/s41593-018-0147-8 (2018).

28 Sutton, R. S. & Barto, A. G. *Reinforcement learning : an introduction*. Second edition. edn, (The MIT Press, 2018).

29 Izawa, J., Criscimagna-Hemminger, S. E. & Shadmehr, R. Cerebellar contributions to reach adaptation and learning sensory consequences of action. *J Neurosci* 32, 4230-4239, doi:10.1523/JNEUROSCI.6353-11.2012 (2012).

30 Ishii, S., Yoshida, W. & Yoshimoto, J. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Netw* 15, 665-687 (2002).

31 Doya, K. Metalearning and neuromodulation. *Neural Netw* 15, 495-506 (2002).

32 Schweighofer, N. & Doya, K. Meta-learning in reinforcement learning. *Neural Netw* 16, 5-9 (2003).

33 Gonzalez Castro, L. N., Hadjiosif, A. M., Hemphill, M. A. & Smith, M. A. Environmental consistency determines the rate of motor adaptation. *Curr Biol* 24, 1050-1061, doi:10.1016/j.cub.2014.03.049 (2014).

34 Hajcak, G. & Foti, D. Errors are aversive: defensive motivation and the error-related negativity. *Psychol Sci* 19, 103-108, doi:10.1111/j.1467-9280.2008.02053.x (2008).

35 Torrecillos, F., Albouy, P., Brochier, T. & Malfait, N. Does the processing of sensory and reward-prediction errors involve common neural resources? Evidence from a frontocentral negative potential modulated by movement execution errors. *J Neurosci* 34, 4845-4856, doi:10.1523/JNEUROSCI.4390-13.2014 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.19.912048; this version posted January 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

36 Kawato, M. & Wolpert, D. Internal models for motor control. *Novartis Found Symp* 218, 291-304; discussion 304-297 (1998).

37 Mathis, M. W., Mathis, A. & Uchida, N. Somatosensory Cortex Plays an Essential Role in Forelimb Motor Adaptation in Mice. *Neuron* 93, 1493-1503 e1496, doi:10.1016/j.neuron.2017.02.049 (2017).

38 Wolpert, D. M., Miall, R. C. & Kawato, M. Internal models in the cerebellum. *Trends Cogn Sci* 2, 338-347 (1998).

39 Ishikawa, T., Tomatsu, S., Izawa, J. & Kakei, S. The cerebro-cerebellum: Could it be loci of forward models? *Neurosci Res* 104, 72-79, doi:10.1016/j.neures.2015.12.003 (2016).

40 Bostan, A. C. & Strick, P. L. The basal ganglia and the cerebellum: nodes in an integrated network. *Nat Rev Neurosci* 19, 338-350, doi:10.1038/s41583-018-0002-7 (2018).

41 Wagner, M. J., Kim, T. H., Savall, J., Schnitzer, M. J. & Luo, L. Cerebellar granule cells encode the expectation of reward. *Nature* 544, 96-100, doi:10.1038/nature21726 (2017).

42 Kostadinov, D., Beau, M., Pozo, M. B. & Hausser, M. Predictive and reactive reward signals conveyed by climbing fiber inputs to cerebellar Purkinje cells. *Nat Neurosci*, doi:10.1038/s41593-019-0381-8 (2019).

43 Verstynen, T. & Sabes, P. N. How each movement changes the next: an experimental and theoretical study of fast adaptive priors in reaching. *J Neurosci* 31, 10050-10059, doi:10.1523/JNEUROSCI.6525-10.2011 (2011).

44 Schaefer, S. Y., Shelly, I. L. & Thoroughman, K. A. Beside the point: motor adaptation without feedback-based error correction in task-irrelevant conditions. *J Neurophysiol* 107, 1247-1256, doi:10.1152/jn.00273.2011 (2012).

45 Neville, K. M. & Cressman, E. K. The influence of awareness on explicit and implicit contributions to visuomotor adaptation over time. *Exp Brain Res* 236, 2047-2059, doi:10.1007/s00221-018-5282-7 (2018).

46 Marko, M. K., Haith, A. M., Harran, M. D. & Shadmehr, R. Sensitivity to prediction error in reach adaptation. *J Neurophysiol* 108, 1752-1763, doi:10.1152/jn.00177.2012 (2012).

47 Taylor, J. A., Klemfuss, N. M. & Ivry, R. B. An explicit strategy prevails when the cerebellum fails to compute movement errors. *Cerebellum* 9, 580-586, doi:10.1007/s12311-010-0201-x (2010).

48 Bates, D., Machler, M., Bolker, B. M. & Walker, S. C. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 67, 1-48 (2015).

49 Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw* 82, 1-26 (2017).

50 Estimated Marginal Means, aka Least-Squares Means (https://CRAN.R-project.org/package=emmeans, 2019).