Interactions between motor exploration and reinforcement learning

Shintaro Uehara1,2, Firas Mawase1, Amanda S. Therrien3,4, Kendra M. Cherry-Allen1, and Pablo Celnik1,3

1Department of Physical Medicine and Rehabilitation, Johns Hopkins Medical Institution, Baltimore, MD, USA; 2Japan Society for the Promotion of Science, Tokyo, Japan; 3Department of Neuroscience, Johns Hopkins Medical Institution, Baltimore, MD, USA; 4Center for Movement Studies, The Kennedy Krieger Institute, Baltimore, MD, USA.

Running head: Increased exploration via reinforcement learning

Correspondence:

Pablo Celnik, MD.
Department of Physical Medicine and Rehabilitation, Johns Hopkins Medical Institution,
600 N. Wolfe St., Baltimore, MD 21287
Email: pcelnik@jhmi.edu

Author contribution: S.U., F.M., A.S.T., and P.C. designed research; S.U., F.M., K.C.A. performed research; S.U. and F.M. analyzed data; S.U. and P.C. wrote the paper.

Number of pages: 27
Number of figures: 6
Number of tables: 1
Number of words: Abstract (250), Introduction (599), Discussion (1631)
Abstract (250 words)

Motor exploration, a trial-and-error process in search for better motor outcomes, is known to serve a critical role in motor learning. This is particularly relevant during reinforcement learning where actions leading to a successful outcome are reinforced while unsuccessful actions are avoided. Although early on motor exploration is beneficial to find the correct solution, maintaining high levels of exploration later in the learning process might be deleterious. Whether and how the level of exploration changes over the course of reinforcement learning, however, remains poorly understood. Here, we evaluated temporal changes in motor exploration while healthy participants learned a reinforcement-based motor task. We defined exploration as the magnitude of trial-to-trial change in movements as a function of whether the preceding trial resulted in success or failure. Participants were required to find the optimal finger-pointing direction using binary feedback of success or failure. We found that the magnitude of exploration gradually increased over time when learning the task. Conversely, exploration remained low in participants who were unable to correctly adjust their pointing direction. Interestingly, exploration remained elevated when participants underwent a second training session, which was associated with faster relearning. These results indicate that the motor system may flexibly upregulate the extent of exploration during reinforcement learning as if acquiring a specific strategy to facilitate subsequent learning. Also, our findings showed that exploration affects reinforcement learning and vice versa, indicating an interactive relationship between them. Reinforcement-based tasks could be used as primers to increase exploratory behavior leading to more efficient subsequent learning.

New & Noteworthy (75 words)

Motor exploration, the ability to search for the correct actions, is critical to learning motor skills. Despite it, whether and how the level of exploration changes over the course of training remains poorly understood. We showed that exploration increased and remained high throughout training of a reinforcement-based motor task. Interestingly, elevated exploration persisted and facilitated subsequent learning. These results suggest that the motor system upregulates exploration as if learning a strategy to facilitate subsequent learning.

Key words: Motor exploration, Trial-and-error, Reinforcement learning, Savings, Meta-learning
Introduction

When learning new motor behaviors, such as when trying a new sport or relearning daily activities after neurological injury, trial-and-error plays a pivotal role. This process, known as motor exploration, helps the motor system to identify the consequences of various actions and update their values based on each movement outcome. This, in turn, allows the person to regulate the expression of the probed actions (Dhawale et al. 2017). Motor exploration has been linked to reinforcement learning (Dhawale et al. 2017), where actions leading to favorable outcomes (i.e., reward) are reinforced and become more frequently expressed, while those resulting in unfavorable outcomes are avoided (Sutton and Barto 1998).

Previous animal and human studies have investigated the impact of motor exploration on motor learning outcomes. Courtship vocalization studies in songbirds indicate that rendition-to-rendition variability in the pitch of their vocalizations, thought to partly serve as motor exploration, supports continuous learning and further optimization of vocalization performance (Fiete et al. 2007; Tumer and Brainard 2007). Similarly, human behavioral studies demonstrated that the magnitude of movement variability or motor exploration is associated with learning rate in a reinforcement-based arm-reaching task that required movement modification toward an optimal pattern (Chen et al. 2017; Therrien et al. 2016; Wu et al. 2014).

While exploration is critical for reinforcement learning, it is not completely understood whether and how the amount of motor exploration changes over the course of reinforcement learning. Learning-related modification in motor exploration and its underlying neural circuit mechanisms were investigated in songbirds studies. Here, as learning proceeds exploratory variability in song production decreases, a change associated to shifts in the control of motor program away from the forebrain to descending motor pathways for vocal control (Aronov et al., 2008) via synaptic reorganization (Garst-Orozco et al., 2014). Much less is known, however, about the interactions between exploration and reinforcement learning in humans. Considering the contribution of exploration on reinforcement learning, it is conceivable that the amount of exploratory behavior is elevated in the early stages of learning when action values are still unknown, but then reduces as one finds a motor pattern that more reliably leads to reward (i.e., optimal motor solution). Alternatively, it is possible that the amount of motor exploration remains elevated throughout training. This could occur if the motor system learns not only the specific task but also acquires knowledge to be exploratory, allowing it to quickly find an optimal solution when exposed to similar conditions in the future (e.g., learning-to-learn, (Braun et al. 2008)).
In the present study, we investigated temporal changes in the amount of motor exploration while healthy participants trained on a goal-directed finger-pointing task. The task was designed to follow a reinforcement learning paradigm in which participants were required to adjust their pointing movement in a trial-by-trial manner toward a predetermined target direction (unbeknown to participants) relying solely on binary feedback about performance outcome (Izawa and Shadmehr 2011; Pekny et al. 2015; Therrien et al. 2016; Uehara et al. 2018). We defined exploration as the magnitude of trial-to-trial change in movement direction as a function of whether the preceding trial resulted in success or failure.

We first studied whether and how the magnitude of motor exploration changes over the course of training on the task (Experiment 1). Secondly, we assessed the impact that modulation of motor exploration may have on subsequent training on the same task (Experiment 2) or on a task that required a different motor solution (Experiment 3). Finally, we reassessed how the magnitude of motor exploration changes over the course of longer task training (Experiment 4).

Materials and Methods

Participants

The study was approved by the Johns Hopkins University School of Medicine Institutional Review Board and was in accordance with the Declaration of Helsinki. A total of 59 healthy participants (25.6 ± 6.3 years, including 35 females, 53 right-handers (self-reported), means ± standard of deviation (SD)) were recruited for the study. All individuals were naïve to the purpose of the study. They provided written informed consent before participating in the study. None of the participants had a history of neurological disease and/or psychological disorders.

Finger-pointing task

Participants performed a center-out finger-pointing task, moving a visually displayed cursor from a central starting location through a target in a shooting movement (Fig. 1a). Participants sat approximately 45 cm in front of a vertical computer monitor (1280 × 1024-pixel resolution). They were instructed to move a digitizing stylus attached on the ventral surface of their index finger on the dominant hand over a horizontal digitizing tablet (48.8 × 30.5 cm active area, Intuos4 XL; Wacom, Saitama, Japan) located on a table. Thus, participants mainly moved the metacarpophalangeal joint of the index finger to control the stylus movement. To facilitate the motions while relaxing the hand, we
asked participants to rest their forearm on an arm support sling that allowed free movements of the arm while eliminating the need for gravitational support. The tablet and participants’ forearm were covered by a box to prevent participants from directly looking at their hand while performing the task. The position of the stylus, sampled at 60 Hz through a custom MATLAB program (R2015b; MathWorks, MA, USA), corresponded to the position of a yellow 1.5-mm-diameter cursor displayed on a black screen such that moving the stylus forward moved the cursor upward. The mapping between the stylus and the displayed cursor displacement (mm) was set as 1:2.

In the task, participants attempted to rapidly move the displayed cursor from a white 3-mm-square starting position centered in the middle of the screen toward a white 16-mm-diameter target in a straight line with no corrections (Fig. 1b). The visible target was always displayed at 90°, 10 cm superior to the starting position. The trials began when the cursor was held in the starting position for 500 ms, leading to the appearance of the target on the screen. Upon presentation of the target, participants started to move their finger so that the cursor crossed through the target. However, the cursor disappeared immediately after participants moved out of the starting position (> 0.3 mm). In other words, participants did not receive online cursor feedback. Instead, reinforcing binary color feedback (green: success, red: failure) was presented to participants at the moment when the invisible cursor passed through the 10-cm radius boundary circle centered around the starting position (i.e. the movement endpoint). The target’s color turned green if the endpoint was within an invisible “success range” (e.g., between the target’s bounds) or red if it missed the range. If a participant’s movement was too fast (< 100 ms) or too slow (> 300 ms), a high- or low pitched auditory tone was provided. Thus, the task was designed so that the movements were not ballistic, but constrained to be executed within a predefined time period. We verbally instructed participants to prioritize task success and then, if possible, to complete the movement in the allowed time window (Table 1). After each trial, participants moved back to the starting position guided by a yellow ring that indicated the distance of the current cursor position from the central starting position. When the invisible cursor was within 1.25 cm from the central starting position, the ring was transformed into the visible cursor.

**Experimental procedure**

**Experiment 1**

We examined whether and how the magnitude of motor exploration changed during training on the reinforcement-based learning task. Twenty participants (25.1 ± 5.7 years; including 14 females)
performed the finger-pointing task in two consecutive blocks (Fig. 1c). The first block started with a 40 trial (one epoch) “baseline” phase (Base1) in which the invisible success range was kept constant between the visible target’s bounds (90° ± 4.5° on the screen). This was followed by a 320 trial (40 trials × eight epochs) “perturbation” phase (Perturb1) where the success range was gradually perturbed from the original range, unbeknown to participants. This was intended to have participants gradually learn a new pointing direction toward a predetermined target angle, set at 110° on the screen (20° counterclockwise rotation from the visible target) through a trial-and-error process (Therrien et al. 2016; Uehara et al. 2018). For this purpose, the left bound of the success range was rotated to 114.5°, whereas the right bound gradually shifted in a trial-by-trial manner according to a moving average of 10 previous pointing directions (i.e., in a “closed-loop” reinforcement schedule). Therefore, to obtain positive feedback, participants were basically required to adjust their pointing direction toward counterclockwise rotation beyond the average of their last 10 trials. When the moving average of 10 last trials fell within the range of 105.5° and 114.5° (the target angle ± 4.5°), this range became the success range so that pointing direction could stay around the target angle. If the moving average exceeded 114.5°, the success range was set between 105.5° and the angle of the moving average to lead pointing direction back to the target angle.

After taking a few minutes break, the participants proceeded to the second block consisting of a 100 trial “post-perturbation” phase (Post). During this phase, the target always turned black irrespective of pointing angles (i.e., no-feedback trials). Before starting the Post phase, we displayed written instructions to participants to repeat the same movements as previously done to make the target green. We implemented these trials to investigate whether adjusted pointing movements continued under the condition where no factor encouraging further motor adjustments was presented. To ensure participants maintain their motivation, we presented the percentage of successful trials on the screen at the end of the first block.

Note that the perturbed direction during the Perturb1 phase was set opposite for left-handers, guiding their pointing direction converged around 70° on the screen (20° clockwise rotation from the visible target). For simplicity’s sake, all other protocols are presented and illustrated in a setting for right hand dominant individuals.

Experiment 2
Results in Experiment 1 showed that the magnitude of motor exploration was elevated and remained high throughout the Perturb1 phase. One possible explanation for this finding is that elevated exploration may persist in order to improve the efficiency of learning during subsequent exposures to the same training situation. To test this assumption, we recruited 15 new participants (27.3 ± 6.4 years; including 10 females) and asked them to perform the task in three consecutive task blocks (Fig. 1d). The first block was composed of the 40 trial Base1 phase and the 320 trial Perturb1 phase as in Experiment 1.

In the second block, the 100 trial Post phase occurred first then followed by a 100 trial “washout” phase (Wash). During the Wash phase, the reinforcing binary feedback was presented again so that the participants were able to gradually get their pointing direction back to the baseline level (towards the visible target) under the same learning context as in the Perturb1 phase. For this purpose, the target angle was set at 90° and the success range was adjusted in a trial-by-trial manner on the basis of the closed-loop reinforcement schedule. The third block was composed of a 40 trial second baseline phase (Base2) and a 320 trial second perturbation phase (Perturb2). The perturbation rule for this block was in line with that implemented in the first block. The percentage of successful trials was presented to the participants at the end of each block. A few minutes of breaks were inserted between the blocks.

**Experiment 3**

Results in Experiment 2 showed that participants were able to relearn the same task more quickly in the second training session, in association with increased exploration from the beginning of the second exposure. It is possible, however, that the rapid relearning was not due to increased exploration, but rather to the presence of a directional bias toward the same rotation direction needed in the second training session. Indeed, previous studies suggested that directional bias in movement and/or in cortical motor representation can be formed via repetition of successful movements (Diedrichsen et al. 2010; Huang et al. 2011; Mawase et al. 2017; Verstynen and Sabes 2011). To rule out this confounder, we performed an additional experiment in which participants were required to adjust their pointing movement during the second training session in the opposite direction to that made in the first training.

We recruited a new group of 14 participants (27.8 ± 7.8 years; including eight females) and asked them to perform a task in three consecutive blocks (Fig. 1e). The task setting in the first two blocks was exactly the same as in Experiment 2. The only difference in this experiment was in the third block where the target angle was set at 70° on the screen (20° clockwise rotation from the visible target) during the Perturb2 phase.
**Experiment 4**

In the previous three experiments, we found that the magnitude of motor exploration was gradually elevated and stayed high throughout the first training. Thus, we asked whether the increased exploratory behavior changes when participants repeat the same tasks for longer periods of time. To investigate this, we recruited a new group of 10 participants (21 ± 0.0 years; including 3 females) and asked them to perform the same task as in Experiment 1, but with a double number of trials (Fig.1f). The first block was composed of the 40 trial Base1 phase, followed by the 320 trial Perturb1 phase, and then the second block of another 320 trial Perturb2 phase. The participants were allowed to have 2-3 minutes break between the blocks to prevent fatigue effects.

**Data analysis**

*Task performance*

Task performance was quantified using pointing angle (PA), the angle between the line connecting the starting position to the center of the visible target and the line connecting the starting position to the endpoint. To analyze right- and left-handed dominances together, we flipped the left-handed data to correspond to right-handed participants. We excluded trials in which pointing angle exceeded |60 °| as outliers (< 0.5 % of trials among all the participants). To quantify learning rate during training sessions, we measured “initial deviation” in pointing angle as follows: Initial deviation (ID) = PA_{Perturb} − PA_{Base}, where PA_{Perturb} represents the mean pointing angle of the first epoch in the perturbation phases (trials 41-80) and PA_{Base} represents the mean pointing angle in the preceding baseline phase (trials 1-40). Only for data in Experiment 3, we flipped the sign of the individual initial deviation from the second training session to match that of the first training session. We also assessed the training-related performance in terms of the mean percentage of successful trials for each of 40-trial epochs on which binary feedback was presented.

*Motor exploration*

For participants to learn the task they needed to actively explore new possible movement directions. Especially when the previous trial resulted in a negative outcome, they needed to change movements in search for a better pointing direction that more reliably led to a positive outcome. Therefore, we defined the magnitude of trial-to-trial change in pointing angle after failed trials as a measure of the amount of
motor exploration (Pekny et al. 2015; Sidarta et al. 2016). In addition, since we wanted to capture the true magnitude changes of the movement deviation regardless of direction, we computed the unsigned magnitude of pointing angle changes (|ΔPA|) from trial \( n \) to trial \( n + 1 \) contingent upon trial \( n \) being success (\( S = 1 \)) or failure (\( S = 0 \)). We then calculated the mean of |ΔPA| after successful or after failed trials respectively for each of 40-trial epochs to track changes in the amount of motor exploration over the course of task training.

**Statistical analysis**

To test whether and how the magnitude of motor exploration changes during task training, we performed a repeated-measures analysis of variance (RMANOVA) to |ΔPA| across epochs during the perturbation phases. For Experiment 1, a two-way RMANOVA was performed with a within-participant factor of OUTCOME (success, failure) and EPOCH (8 epochs). For Experiment 2 and 3, a three-way RMANOVA was performed with a within-participant factor of OUTCOME, EPOCH, and SESSION (first and second training sessions).

Only for Experiment 2 and 3, we separately performed a paired-\( t \) test between |ΔPA| of the first epoch in the Perturb1 and that in the Perturb2 to evaluate a priori hypothesis that exploratory behavior would be greater from the beginning of the second than the first training session. A paired-\( t \) test was also used to compare initial deviation (ID) in pointing angle (a proxy for learning rate) between the first and the second training sessions. To assess the relationship between the amount of exploration at the beginning of task training and learning rate, we applied Pearson’s correlation analysis between |ΔPA| of the first epoch in the perturbation phase and initial deviation. This was separately performed for the first and the second training sessions, respectively. Finally, to test whether the changes in the magnitude of exploration from the first to the second training sessions was associated with changes in learning rate between the two sessions, we performed Pearson’s correlation analysis between (|ΔPA|\text{Second} − |ΔPA|\text{First}) and (ID\text{Second} − ID\text{First}), where |ΔPA|\text{First} and |ΔPA|\text{Second} represent |ΔPA| of the first epoch in the Perturb1 and Perturb2 phases, and ID\text{First} and ID\text{Second} represent initial deviation in pointing angle in the first and the second training sessions, respectively.

To determine whether participants maintained the learned pointing direction during the Post phase, we compared the average pointing angle of the first 40-trial epoch in the Post phase with the epoch of the Base1 phase using a paired-\( t \) test. Similarly, to determine whether pointing direction successfully returned to baseline level during the Wash and the subsequent Base2 phases, we compared
the epoch in the Base1 phase with the last 40-trial epoch in the Wash phase and the epoch in the Base2 phase respectively using a paired-\(t\) test.

All statistical analyses were performed using SPSS (version 20; IBM, Armonk, NY). All RMANOVA were tested for the assumption of homogeneity of variance using Mauchly’s Test of Sphericity. For those tests in which this assumption was violated, Greenhouse-Geisser corrections statistic was reported. Effects were considered significant if \(p \leq 0.05\). Effect sizes were reported in Cohen’s \(d_z\) value (\(d_z\)) for paired-\(t\) test, Cohen’s \(d\) value (\(d\)) for unpaired-\(t\) test, and partial eta squared value (\(\eta^2_p\)) for ANOVA, respectively.

**Results**

*Classification of “learners” and “non-learners”*

As observed in our previous study (Uehara et al. 2018), we found a subset of participants who did not adjust their pointing angle sufficiently to reach the target angle in a limited number of perturbed trials. These non-learners were analyzed separately from the rest of the participants (learners). Moreover, they served as a control group for our investigation into the relationship between a modification in motor exploration and task learning, rather than simple task execution. We defined non-learners if the moving average of pointing angles did not exceed \(|18^\circ|\) (90% of the target angle) during the Perturb1 phase (320 trials). With this criterion, five out of 20 (Experiment 1), three out of 15 (Experiment 2), two out of 14 participants (Experiment 3) were classified as non-learners respectively (about 20% of the total participants, but no participants from Experiment 4), and all the non-learners were integrated into one group (Non-learner group, \(n = 10, 26.7 \pm 7.1\) years; including six females). Consequently, 15 (23.9 \(\pm 3.7\) years; including 11 females), 12 (29.0 \(\pm 6.9\) years; including nine females) and 11 (26.7 \(\pm 7.1\) years; including six females) participants from Experiment 1, 2 and 3 were defined as learners and proceeded to the main analyses. Note that one participant from Experiment 3 was excluded as an outlier from the analysis since she/he did not show gradual and systematic learning of the task. Specifically, this participant’s moving average of 10 pointing angles moved back to the baseline level even after reaching the predetermined target angle once (20° counter-clockwise rotation) during the Perturb 1 phase. Furthermore, the participant moved into the clockwise rotation direction beyond the baseline angle.

*Motor exploration increased during training*
In Experiment 1, we investigated whether and how the amount of exploratory behavior changes during training on the reinforcement-based motor task.

We first confirmed that the participants successfully learned the task, indicated by pointing angle shifts on trial-by-trial bases from the original toward the target direction while exposed to the perturbation (Fig. 2). During the Base1, the participants performed the task accurately given that pointing direction was within the visible target (mean pointing angle, $-1.0 \pm 0.7^\circ$, mean $\pm$ SE, Fig. 2a) with high accuracy (percentage of trial success, $70.2 \pm 4.3\%$, Fig. 2b). During the Perturb1, pointing angle gradually shifted toward the predetermined target angle and finally converged on $18.9 \pm 0.8^\circ$ in the last 40-trial epoch (Fig. 2a). This resulted in significant errors early on, accompanied by a slight but gradual increase in the percentage of successful trials across epochs (Fig. 2b). This new movement direction was maintained during the subsequent Post phase (average of the first 40 trials, $13.4 \pm 1.9^\circ$) as compared with the Base1 (paired-t test, $t_{14} = 7.2$, $p < 0.001$, $d_z = 1.85$), indicating retention of the newly learned motor pattern.

As a proxy for motor exploration, we computed the magnitude of trial-to-trial changes in pointing angle ($|\Delta PA|$) as a function of whether the initial trial resulted in success or failure. We found that on average $|\Delta PA|$ after failed trials was greater than that after successful trials (RMANOVA [OUTCOME (2) $\times$ EPOCH (8)], effect for the factor OUTCOME, $F_{1,14} = 52.4$, $p < 0.001$, $\eta^2_p = 0.79$, Fig. 2c). This was also revealed by broader probability distributions of $|\Delta PA|$ for trials after failure than after success (top panels in Fig. 2d). These results indicate that failing to get positive feedback led to greater trial-to-trial movement changes, presumably in search for a better motor solution (Pekny et al. 2015; Sidarta et al. 2016). Importantly, when we focused on temporal changes in $|\Delta PA|$ over the course of training, we found that the magnitude of $|\Delta PA|$ after failed trials was not constant but gradually increased across epochs and remained elevated throughout the Perturb1 phase (Fig. 2c). Though small, similar temporal change was observed in $|\Delta PA|$ after successful trials (effect for the factor EPOCH, $F_{7,98} = 2.7$, $p = 0.01$, $\eta^2_p = 0.16$; OUTCOME $\times$ EPOCH interaction, $F_{7,98} = 1.3$, $p = 0.25$, $\eta^2_p = 0.09$). This increase was qualitatively represented as a broader probability distribution of $|\Delta PA|$ at the last (8th) epoch compared to the first epoch (top panels in Fig. 2d). In a similar vein, we found wider distribution in the signed magnitude of trial-to-trial changes ($\Delta PA$, bottom panels in Fig. 2d) into both positive and negative directions in the last epoch (standard deviation of $\Delta PA$, $9.0 \pm 1.0^\circ$) relative to the first epoch ($5.7 \pm 0.9^\circ$, paired-t test, $t_{15} = 3.0$, $p = 0.01$, $d_z = 0.76$). Although there were small biases of the movement-correcting direction in the counter-clockwise direction particularly after failed trials (bottom...
panels in Fig. 2d), the increase in |\(\Delta PA\)| can be regarded as increased exploratory behavior in both counter-clockwise and clockwise directions rather than in one particular direction.

In sum, these results demonstrate that the amount of motor exploration is gradually up-regulated and remained elevated throughout training on the task.

*Increased exploration continued and facilitated relearning of the same task*

Given that the amount of motor exploration increased and persisted throughout training, in Experiment 2 we tested whether this increase was maintained and benefited subsequent training on the same reinforcement-based task.

First, we replicated the results of Experiment 1 in a separate group of participants; that is training of the task resulted in a gradual increase in exploratory behavior during the first training exposure (Fig. 3). During the Base1, the participants’ pointing direction converged on the visible target (mean pointing angle, 0.5 ± 0.9°, Fig. 3a) with high accuracy (percentage of trial success, 69.9 ± 4.8%, Fig. 3b). When exposed to the Perturb1, the pointing angle gradually shifted (initial deviation, 3.3 ± 0.9°, Fig. 3c) to end up near the target angle of 18.4 ± 1.0° in the last 40-trial epoch (Fig. 3a). This was accompanied by low number of successful trials early on with a gradual increase across epochs (Fig. 3b). The new pointing angle persisted to some degree during the subsequent Post phase (average of the first 40 trials, 9.8 ± 1.7°), remaining significantly greater when compared to Base1 (paired-\(t\) test, \(t_{11} = 5.1, p < 0.001, d_z = 1.46\)). Importantly, as in Experiment 1, we found gradual increase in |\(\Delta PA\)| across epochs during the Perturb1 phase along with task learning (Fig. 3d and 3e). This increase was evident after failed trials, but not after successful trials.

Second, we found that during the second exposure to the same task, the participants were able to adjust their pointing direction more quickly than in the first training session (savings effect). Of note, we confirmed that before the second exposure, participants’ pointing angle returned to the baseline level during the Wash phase (i.e., learned movement pattern was washed out; average of the last 40 trials, 2.1 ± 1.0°; paired-\(t\) test comparing to Base1, \(t_{11} = 1.2, p = 0.25, d_z = 0.35\), Fig. 3a). This pointing angle persisted during the following Base2 phase (0.9 ± 0.7°; \(t_{11} = 0.3, p = 0.79, d_z = 0.08\)) and resulted in high percentage of successful trials (69.0 ± 4.9%, Fig. 3b). When exposed to the Perturb2, however, pointing angle shifted toward the target direction more quickly (initial deviation, 7.8 ± 1.6°) than during the Perturb1 (paired-\(t\) test, \(t_{11} = 2.5, p = 0.03, d_z = 0.71\), Fig. 3c). At the end of training, the participants expressed similar amount of angular changes as in the Perturb1 (19.4 ± 1.4° in the last 40-trial epoch).
Strikingly, $|\Delta PA|$ after failed trials remained increased throughout the Wash phase (last 40-trial epoch in the Perturb1, 7.9 ± 1.3°, Wash, 7.3 ± 1.4°; $t_{11} = 0.3$, $p = 0.7$, $d_z = 0.76$). This was followed by greater error corrections from the onset of the Perturb2 (first epoch, 8.2 ± 1.2°) compared with that in the Perturb1 (4.7 ± 0.7°, paired-\textit{t} test, $t_{11} = 2.6$, $p = 0.02$, $d_z = 0.76$, Fig. 3d). Furthermore, this increased magnitude in $|\Delta PA|$ after failed trials persisted throughout training, unlike the pattern observed during the first training session (\textit{RM}ANOVA [\textit{OUTCOME} (2) × \textit{EPOCH} (8) × \textit{SESSION} (2)], interaction among three factors, $F_{7,77} = 3.4$, $p = 0.003$, $\eta^2_p = 0.24$; \textit{EPOCH} × \textit{SESSION} interaction, $F_{7,77} = 2.2$, $p = 0.045$, $\eta^2_p = 0.17$; effect for the factor \textit{OUTCOME}, $F_{1,11} = 32.6$, $p < 0.001$, $\eta^2_p = 0.75$). Additionally, we found no difference in the magnitude of $|\Delta PA|$ after successful trials at the training onset between the two sessions (Perturb1, 3.9 ± 0.4°, Perturb2, 3.7 ± 0.4°; paired-\textit{t} test, $t_{11} = 0.5$, $p = 0.66$, $d_z = 0.13$).

These changes in exploratory behavior were qualitatively visualized in a broader probability distribution of both unsigned and signed magnitude of trial-to-trial changes (Fig. 3e), albeit small biases of the movement-correcting direction in the direction of the perturbed success range could be observed (bottom panels in Fig. 3e).

To determine potential associations between learning rate and the amount of exploration, we performed a correlation analysis. In both the Perturb1 and the Perturb2 we found a positive correlation between $|\Delta PA|$ after failed trials from the first epoch and initial deviation ($r = 0.64$, $p = 0.03$ and $r = 0.74$, $p = 0.01$, respectively, Fig. 3f). Furthermore, the magnitude of changes in $|\Delta PA|$ after failed trials from the first to the second training sessions was correlated with the difference in learning rate between the two sessions ($r = 0.61$, $p = 0.04$, Fig. 3g). In contrast, we did not find any significant correlations for $|\Delta PA|$ after successful trials (Perturb1, $r = 0.35$, $p = 0.26$; Perturb2, $r = 0.10$, $p = 0.76$; changes from Perturb1 to Perturb2, $r = 0.53$, $p = 0.07$).

In sum, these results demonstrate that elevated exploration during the first training session persists during subsequent bouts of training, and is associated with facilitation of subsequent learning.

**Increased exploration facilitated learning on a second task that required a different motor solution**

Although in Experiment 2 increased exploration was associated with learning facilitation during the second training, it is possible that this effect was simply due to the contribution of a directional bias toward the same rotation direction in the second training. To rule out this confounder, a new group of participants took part in Experiment 3 where the Perturb2 phase required adjusting the pointing direction...
in the opposite direction to that made in the Perturb1 phase. In this manner, faster relearning would be attributed to increased motor exploration rather than a formed directional bias.

Similar to Experiments 1 and 2, the amount of motor exploration showed a gradual increase as the participants learned the task during the first training session (Fig. 4). During the first Base1 phase, the participants’ pointing direction converged on the visible target (pointing angle, 1.5 ± 0.9°, Fig. 4a) with high accuracy (percentage of trial success, 66.8 ± 5.2%, Fig. 4b). Then, when exposed to the Perturb1 phase, the pointing angle gradually shifted (initial deviation, 2.6 ± 1.5°, Fig. 4c) to end up near the target angle of 17.6 ± 1.1° at the last 40-trial epoch, Fig. 4a). This was accompanied by a low number of successful trials early on with a gradual increase across epochs (Fig. 4b). During the subsequent Post phase, the new pointing angle still persisted (average of the first 40 trials, 15.7 ± 2.1°) without returning to the level of the Base1 phase (paired-t test, $t_{10} = 5.6$, $p < 0.001$, $d_z = 1.68$).

Importantly, we replicated our previous findings that $|\Delta PA|$ after failed trials showed a gradual increase across epochs throughout the Perturb1 phase (Fig. 4d).

When exposed to the second training session, the participants were able to learn the task more quickly than in the first training, similar to Experiment 2. We confirmed that the learned pointing angle was washed out, i.e., participants’ pointing angle returned to the level of the Base1 phase during the Wash phase (average of the last 40 trials, 1.5 ± 1.0°; paired-t test, $t_{10} = 0.1$, $p = 0.94$, $d_z = 0.02$, Fig. 4a). This angle direction persisted during the subsequent Base2 phase (1.6 ± 0.6°; paired-t test, $t_{10} = 0.1$, $p = 0.95$, $d_z = 0.02$) and resulted in high percentage of task success (percentage of trial success, 68.0 ± 6.3%, Fig. 4b). Importantly, when exposed to the Perturb2, their pointing angle shifted toward the new target direction (initial deviation, 8.8 ± 1.4°) more quickly than during the Perturb1 (paired-t test, $t_{10} = 4.0$, $p = 0.002$, $d_z = 1.22$, Fig. 4c) even when the second training session required adjusting the pointing direction in the opposite direction (last 40-trial epoch, −18.6 ± 1.0°) to that in the first training. Moreover, accompanying this faster learning, $|\Delta PA|$ after failed trials remained increased during the Wash phase (last 40-trial epoch in the Perturb1, 6.1 ± 0.4°, Wash, 6.2 ± 1.0°; $t_{10} = 0.1$, $p = 0.9$, $d_z = 0.03$) and showed greater magnitude from the onset of the second training compared to the first training (the first epoch of the Perturb1, 4.0 ± 0.6 °, Perturb2, 5.8 ± 0.8°; paired-t test, $t_{10} = 2.4$, $p = 0.04$, $d_z = 0.71$, Fig. 4d).

However, unlike Experiment 2, $|\Delta PA|$ after failed trials continued to increase through the second training ($\text{RMANOVA } \left[\text{OUTCOME } (2) \times \text{EPOCH } (8) \times \text{SESSION } (2)\right]$, effect for the factor SESSION, $F_{1,10} = 10.4$, $p = 0.009$, $\eta_{p}^2 = 0.51$). This temporal pattern was different from that of $|\Delta PA|$ after success (OUTCOME × SESSION interaction, $F_{1,10} = 11.8$, $p = 0.006$, $\eta_{p}^2 = 0.54$; a factor OUTCOME, $F_{1,10} =
41.5, $p < 0.001$, $\eta_p^2 = 0.81$) in which we found no difference in the magnitude at the onset between the two training sessions (the first epoch of the Perturb1, $2.5 \pm 0.2^\circ$, Perturb2, $3.0 \pm 0.3^\circ$; paired-$t$ test, $t_{10} = 1.3$, $p = 0.22$, $d_z = 0.40$). Similar to the previous two experiments, changes in exploratory behavior were qualitatively visualized in a broader probability distribution of the magnitude of trial-to-trial changes regardless of small biases of the movement-correcting direction (Fig. 4e).

Consistent with Experiment 2 findings, we observed a positive correlation between $|\Delta PA|$ after failed trials of the first epoch and initial deviation in both the Perturb1 and the Perturb2 ($r = 0.87$, $p = 0.001$ and $r = 0.61$, $p = 0.04$, respectively, Fig. 4f). Furthermore, changes in $|\Delta PA|$ after failed trials from the first to the second training sessions were correlated with changes in learning rate between the two sessions ($r = 0.67$, $p = 0.03$, Fig. 4g). We did not find any significant correlations between $|\Delta PA|$ after successful trials and learning rate (Perturb1, $r = 0.45$, $p = 0.16$; Perturb2, $r = 0.36$, $p = 0.27$; changes from Perturb1 to Perturb2, $r = 0.44$, $p = 0.18$).

Although we observed faster learning in the second training session, there still remains a possibility that the second task requiring movement adjustment toward clockwise rotation direction might be by default easier to learn as compared with the first task. To rule out this possibility, we reanalyzed data from our previous work (Uehara et al., 2018) in which healthy participants ($n = 12$) performed the same task without a prior training session. The only difference was that in the previous study the target angle was set at $30^\circ$, instead of $20^\circ$, clockwise rotation from the visible target. The initial deviation in pointing angle showed comparable magnitude ($4.9 \pm 1.2^\circ$) to that of the first task in the present study (unpaired-$t$ test, $t_{25} = 0.5$, $p = 0.62$, $d = 0.20$; $t_{22} = 1.1$, $p = 0.28$, $d = 0.45$; $t_{21} = 0.8$, $p = 0.24$, $d = 0.50$, Experiment 1, 2 and 3, respectively). These results indicate that the faster learning during the second training session is not due to the specific training angle direction.

Finally, to determine whether the small directional bias plays any effects on learning the second task, we directly compared the initial deviation in the Perturb2 phase between Experiment 2 and 3 using an unpaired-$t$ test. The result revealed no significant differences between them ($t_{21} = 0.5$, $p = 0.64$, $d = 0.20$), suggesting there was no clear benefit of the small directional bias to the learning facilitation in the second training session.

Overall Experiment 3 revealed similar results as in Experiment 2; that is, faster learning of the second task and its association with the increased exploratory behavior. These results indicate that faster relearning in the second training is not largely attributed to a potential directional bias and retrieval of the original memory.
Increased exploration continued even after longer training of the same task

In the prior experiments, we found that participants’ motor exploration increased and remained elevated throughout the first task training. Thus, we asked whether the magnitude changes with longer repetition of trials.

As found in the previous three experiments, participants’ pointing angle gradually shifted toward the predetermined target angle and converged on $17.2 \pm 1.5^\circ$ in the last 40-trial epoch during the Perturb1 phase (Fig. 5a). This was accompanied by a gradual increase in the percentage of successful trials across epochs (Fig. 5b). During the subsequent Perturb2 phase, the shifted pointing angle slightly improved and remain at the target until the end of the phase (last 40-trial epoch, $19.4 \pm 0.7^\circ$). As found in the prior experiments, the magnitude of $|\Delta PA|$ after failed trials showed gradual increases (Fig. 5c). Interestingly, this changed remained high during Perturb2 phase (first and last 40 trial epochs, $6.4 \pm 1.1^\circ$, $6.4 \pm 0.8^\circ$), despite the continuous trial repetition in the same task. These results indicate that the magnitude of motor exploration persisted elevated even during longer training of the same task. Of note, the percentage of task success trials did not show a clear return to baseline levels ($78.3 \pm 4.5 \%$), even at the end of the Perturb2 phase (last epoch, $64.8 \pm 4.9 \%$, paired- $t$ test, $t_9 = 2.2$, $p = 0.055$, $d_z = 0.70$).

Non-learners showed little exploratory behavior throughout training

We analyzed the Non-learner group separately. This group constitutes the ideal control group to investigate whether the upregulated motor exploration during task training is associated with learning rather than simple task execution.

Task performance of the Non-learner group during the Baseline phase was similar to that of the learners from the three experiments. Indeed, their pointing angle converged on the visible target (pointing angle, $1.0 \pm 0.9^\circ$, Fig. 6a) and resulted in high task accuracy (percentage of trial success, $70.3 \pm 4.8 \%$, Fig. 6b). However, their pointing angle did not show particular shifts toward the target angle when exposed to the Perturb1 phase, ending up with $5.8 \pm 1.3^\circ$ in the last 40-trial epoch (Fig. 6a). This was accompanied by no increase, but rather a decrease, in task accuracy throughout training (Fig. 6b). Additionally, unlike the learners, we did not find a gradual increase in the magnitude of $|\Delta PA|$ either after failed or successful trials throughout training ($rM$ANOVA [OUTCOME (2) × EPOCH (8)], OUTCOME × EPOCH interaction, $F_{1.1, 10.2} = 0.9$, $p = 0.39$, $\eta_p^2 = 0.10$, effect for the factor EPOCH, $F_{1.2}$,


This result supports the view that, rather than mere task execution, the learners’ gradual increase in exploratory behavior during task training was largely attributed to learning. Furthermore, in the Non-learner group, we found a comparable magnitude of $|\Delta PA|$ after failed trials to that after successful trials (effect for the factor OUTCOME, $F_{1,9} = 4.6, p = 0.06, \eta_p^2 = 0.34$). This result suggests that the Non-learner group may have less sensitivity to the negative feedback than the learners.

To further investigate feedback sensitivity of the Non-learner group and compared them with the learners, we extended trial-to-trial analysis for the magnitude of $|\Delta PA|$ to include the history of past feedback (Pekny et al. 2015). Here, we considered all possible eight combinations of success and failure feedback for three consecutive trials. The feedback history for three consecutive trials was represented by variables $S(n)$, $S(n-1)$, and $S(n-2)$, indicating whether task performance was successful in trials $n$, $n-1$, and $n-2$, respectively (Fig. 6d). For this analysis, we used trials only from the Base1 and Perturb1 phases that were completed by all the participants.

When we first analyzed $|\Delta PA|$ as a function of the feedback in the learners, we did not find significant differences among the participants across the four experiments (a two-way mixed-effect RMANOVA [HISTORY (8 combinations) × EXPERIMENT (Experiment 1, 2, 3, and 4)], effect for the factor EXPERIMENT ($F_{3,44} = 1.3, p = 0.29, \eta_p^2 = 0.08$), HISTORY × EXPERIMENT interaction ($F_{4.5, 65.7} = 0.4, p = 0.83, \eta_p^2 = 0.03$). Therefore, we grouped all the learners from the three experiments into one group (Learner group, $n = 38$) and compared them with the Non-learners. We found that $|\Delta PA|$ gradually increased as unsuccessful feedback history accumulated in the Learner group, whereas in the Non-learner group it showed little increase in response to failure feedback accumulation (RMANOVA [HISTORY (8) × GROUP (Learner, Non-learner)], HISTORY × GROUP interaction, $F_{1.6, 60.2} = 6.4, p = 0.005, \eta_p^2 = 0.10$, Fig. 6d). This result indicates that the non-learners have less sensitivity to unsuccessful feedback, which may lead to less exploratory behavior and learning.

**Discussion**

Exploring for the correct motor actions is a critical element of reinforcement motor learning. However, whether and how the exploratory behavior changes as individuals learn a new task has not been well characterized. As predicted, we found that participants’ motor exploration gradually increased as they trained on a reinforcement-based motor task. However, the exploratory behavior remained elevated even after clear improvements, and stability, in task execution accuracy. Furthermore, participants showed
sustained increased motor exploration when they were exposed to a second bout of training in the same task. This effect was associated and proportional to faster relearning. In contrast, participants who were unable to sufficiently learn the task demonstrated few changes in exploratory behavior. This finding confirmed that exploration is critical to reinforcement learning and indicates that the gradual increase in motor exploration is not the result of mere task execution. In addition, our findings suggest that the motor system can upregulate the amount of motor exploration during learning a reinforcement-based motor task as if acquiring a beneficial strategy that facilitates subsequent learning.

In the motor domain, reinforcement learning has been described as the process in which actions leading to successful outcomes are reinforced while those leading to unsuccessful outcomes are avoided (Sutton and Barto 1998). This form of learning necessitates exploratory behavior through which the motor system identifies or updates values of potential actions in trial-by-trial manner based on each action outcome. Therefore, as observed here, an increase in motor exploration should be expected during reinforcement learning so that the motor system can identify the reward landscape in action space. However, previously it was not clear whether the increased exploratory behavior returns to baseline levels once the task has been learned (consistent high level of performance success).

In this study, we found that the amount of exploration remained elevated throughout training. This was observed even when the percentage of successful trials continued to increase throughout training. At first glance, this result might seem contradictory to previous findings indicating that the amount of exploratory behavior can be regulated by an overall rewarding situation, e.g. the probability of reward. For instance, studies across a variety of species indicated that the overall trial-to-trial change in motor output tend to decrease under a condition where the probability for positive reward outcome increases (Galea et al. 2013; Gharib et al. 2001; Gharib et al. 2004; Pekny et al. 2015; Stahlman et al. 2010; Stahlman and Blaisdell 2011; Takikawa et al. 2002). There are some important differences between those studies and ours. First, we measured trial-to-trial movement changes after failed trials as a proxy for exploratory behavior, whereas previous investigations measured trial-to-trial changes regardless of the previous outcome (i.e. overall variability). Second, we tracked changes in exploration as a function of learning-associated changes in reward, whereas previously overall variability was determined in separate blocks across a variety of experimentally-controlled rewarding conditions. Nevertheless, these previous findings showing the overall reduction on movement-to-movement variability do not contradict that changes after failure increase in magnitude and remain elevated in the presence of errors.
We found that the exploratory behavior started at a greater magnitude from the onset of the second training when compared to the first training, even when the movement directions were different. This initial increase in exploration was associated, and proportional, to faster learning of the second task relative to the first task. These results can be interpreted as the motor system learns not only a new task-specific motor pattern, but also gains the strategy of being more exploratory. In other words, it acquires knowledge how to learn different tasks given the same context (e.g., learning-to-learn or meta-learning, (Braun et al. 2009; Braun et al. 2010; Krakauer and Mazzoni 2011)) to increase the efficiency of subsequent training. However, we cannot conclude whether the motor system actually “learned” the strategy or remained in a heightened exploratory state since the percentage of task success remained low. To determine this, in Experiment 4 participants repeated more training trials, to evaluate time-course changes in the magnitude of exploration while the success rate becomes substantially higher. We found, however, that the rate of task success still remained lower than baseline even after longer repetition of trials. Thus, we were unable to dissociate the effects leading to a heightened exploration; a lower success rate may have kept motor exploration greater, or alternatively greater exploration (i.e. greater magnitude of movement changes after failure) as a learned strategy may have resulted in a certain probability of task failure (i.e. lower success rate). Future studies should consider experimental designs to disentangle this relationship.

The present study cannot clarify if the increased exploratory behavior represents “explicit” strategic changes that participants intentionally controlled. Interestingly, our results showed different patterns of motor performance during the Post phase (i.e. no-feedback trials) among the three experiments, such as a distinct drop relative to the end of the Perturb1 phase and/or an upward drift (Fig. 2a, 3a and 4a). Although these results were somewhat unexpected because all the participants experienced exactly the same training schedule until the end of the Post phase, these seemingly contradictory results may be due to a different degree of dependency on explicit components across individuals during task training or during the subsequent Post phase. Indeed, a recent human behavioral study suggests that explicit strategy is partly engaged in the process of reinforcement learning, since motor patterns learned through reinforcement mechanisms were degraded when the use of explicit strategy was experimentally constrained or intentionally removed (Holland et al., 2018). On the other hand, generating identical movements in successive trials is virtually impossible. In other words, movements exhibit trial-to-trial variability regardless of outcomes in a preceding trial. This type of movement variability or “motor noise” is thought to be, in part, an inherent feature originated from the
fundamental properties of the neuro-muscular system (Dhawale et al. 2017; Faisal et al. 2008; Jones et al. 2002; van Beers et al. 2004). Therefore, movement changes in response to failed outcomes include motor noise-related changes as well as exploratory action changes. We think the contribution of motor noise to the gradual increase in motor exploration after failed trials during training, however, is very limited. This is because motor noise is also present after successful trials and in the non-learner group, yet these two situations did not lead to changes in exploratory behavior. Thus, we interpret the increase in exploration after failed trials during learning largely related to an exploratory strategy, irrespective of the level of awareness, rather than simple motor noise.

The increased exploratory behavior led to faster learning in subsequent practice exposures, a phenomenon known as savings (Krakauer et al. 2005; Mawase et al. 2014; Mawase et al. 2017; Zarahn et al. 2008). Although debate continues as to which mechanisms contribute to savings (Huang et al. 2011; Leow et al. 2016; Morehead et al. 2015; Orban de Xivry and Lefevre 2015; Roemmich and Bastian 2015), a study suggested the potential contribution of movement direction bias resulting from successful movement repetitions (Huang et al. 2011). In addition, it is possible that the observed small biases in movement correction direction may have also contributed to faster learning. Although, this hypothesis cannot explain why participants experienced faster relearning in Experiment 3 where the biases were in the opposite direction to the appropriate movement corrections. The faster relearning in Experiment 3 also cannot be attributed to prior exposure to the Wash phase, where participants experienced movements in the same rotation direction as in the second training. This is because any tangible effect should have been then present in Experiment 2, where participants experienced a washout in the opposite direction to the second training, yet no negative effect of bias was observed in the second training exposure. Therefore, direction biases cannot fully explain the faster relearning observed in the second training.

Our Experiment 3 results further posit that the effect of increased motor exploration can be generalized to another task which requires a different motor solution but relies on the same reinforcement learning paradigm. Therefore, it is conceivable that training on reinforcement-based tasks could be used as a primer to increase exploratory behavior and enhance the efficiency of subsequent learning in the same contextual setting. It could be argued that our findings are specific to the nature of our current experimental paradigm testing reinforcement learning. Here, we used a task in which the target direction cannot be learned in any other ways except through exploration and reinforcing feedback. Therefore, the present findings might only be applicable to reinforcement forms of learning.
Generalization to other learning forms (e.g. error-based learning, use-dependent learning) should be done with caution and formally tested in future studies.

Interestingly, some participants were unable to reach the optimal motor pattern during the first training session. These individuals did not show a gradual increase in exploratory behavior throughout training. This group of participants seems to have less sensitivity to binary feedback, particularly to negative feedback. Although it cannot be distinguished whether the reduced susceptibility is accounted for by the level of perception or the level of motor planning integration, the finding can explain why the non-learners exhibited less learning efficiency and therefore no gradual increase in exploration during training.

In summary, the present study demonstrated that the motor system upregulates the amount of motor exploration during reinforcement learning. In turn, this effect is associated with improvements in learning efficiency during subsequent training in the same contextual setting. Although we cannot be conclusive, our findings might suggest that the motor system acquires knowledge to become more exploratory as if developing an overall strategy that can be useful to learn new motor actions, at least in the presence of similar context. The present results open an opportunity to design better training paradigms for healthy individuals as well as for people undergoing rehabilitation to speed up learning.

Acknowledgments
We thank Tziporah Thompson for illustration on Fig. 1a.

Grants
This work was supported by NIH 4R01HD073147-05 (to P.C) and JSPS Overseas Research Fellowship (to S.U).

Conflict of interest
The authors declare no competing financial interests
Reference


**Figure legends**

**Fig 1. Experimental protocols.**

a. Experimental setup. Participants moved a stylus on a digitizing tablet with their index finger. b. Finger-pointing task. Participants performed a pointing movement from a central starting position to around a visible white target. Binary feedback (target color) about task performance was presented instead of vector cursor feedback. Participants were instructed to obtain positive (green) color feedback at every trial. c. Experiment 1. Participants performed the task in two blocks, composed of baseline (Base1), perturbation (Perturb1), and post-perturbation phases (Post). During the Perturb1, a range for success feedback (gray area) gradually shifted from the original range towards counter-clockwise direction according to a moving average of the previous 10 pointing direction (blue line). No-feedback trials were implemented for the Post phase. d. Experiment 2. Participants performed the task in three blocks, composed of Base1, Perturb1, Post, Washout (Wash), Base2, and Perturb2. During the Wash phase, the binary feedback was presented so that pointing direction returned to the baseline level. The task setting in the following Base2 and Perturb2 was exactly the same as in the Base1 and Perturb1. e. Experiment 3. Participants performed the task in a similar setting as in Experiment 2. The only difference was that the target angle was set at clockwise direction during the Perturb2. The numbers under each phase in the x-axis represent the amount of trials. The vertical axis represents rotation angle from the center of the visible target. Positive value indicates counter-clockwise rotation. Note that the direction of the perturbation was set to the opposite side for left-handed participants. f. Experiment 4. Participants performed the task in a similar setting as in Experiment 1 but with more number of trials for the Perturb phase.

**Fig 2. Gradual increase in exploratory behavior during task training.**

a. Pointing angle in the baseline (Base1), perturbation (Perturb1), and post-perturbation (Post) phases. Positive value indicates counter-clockwise direction relative to the target position. Dots and shaded area show the mean and standard errors of the mean (SEM) for each trial. Solid vertical line indicates time break between the blocks. b. Mean percentage of successful trials for each 40-trial epoch. c. Mean trial-to-trial unsigned changes in pointing angle (|ΔPA|) for each epoch. Blue and gray open dots indicate trials after failure (S = 0) and success (S = 1) respectively. d. Probability distribution of unsigned (|ΔPA|, top panels) and signed (ΔPA, bottom panels) trial-to-trial changes in pointing angle (S = 0 in blue, S = 1 in gray) from the first (left panels) and the last (right panels) epochs during the Perturb1 phase. The fit of unsigned or signed changes to a folded normal or a normal distribution are plotted over the histogram. Data from all
the participants are pooled together on this analysis. Values inside the bottom panels show the mean and SEM of $\Delta PA$ across participants.

**Fig 3. Persistence of increased exploration and faster relearning during exposed to the second training of the same task.**

- **a.** Pointing angle while training the task. Dots and shaded area show the mean and SEM for each trial in the Base1-Perturb1 (blue), Post-Wash (black), and Base2-Perturb2 (red) phases. Solid vertical lines indicate time break between the blocks.
- **b.** Mean percentage of successful trials for each epoch.
- **c.** Initial deviation in pointing angle for the Perturb1 (blue) and Perturb2 (red).
- **d.** $|\Delta PA|$ across epochs. Blue and red dots indicate trials after failure ($S = 0$) in the first and the second training sessions, respectively. Gray dots indicate trials after success ($S = 1$).
- **e.** Probability distribution of unsigned ($|\Delta PA|$, top panels) and signed ($\Delta PA$, bottom panels) trial-to-trial changes in pointing angle from the first and the last epochs during the Perturb1 and Perturb2 phases. The fit of unsigned or signed changes to a folded normal or a normal distribution are plotted over the histogram. Dashed lines indicate regression line. Values inside the bottom panels show the mean and SEM of $\Delta PA$ across participants.
- **f.** Relationship between $|\Delta PA|$ ($S = 0$, first epoch) and initial deviation for the Perturb1 (blue) and the Perturb2 (red).
- **g.** Relationship between the magnitude of changes in $|\Delta PA|$ ($S = 0$, first epoch) and changes in initial deviation from the first to the second training sessions.

**Fig 4. Increased exploration facilitated learning in the second task requiring a different motor solution.**

- **a.** Pointing angle while training the task. Dots and shaded area show the mean and SEM for each trial during the Base1-Perturb1 (blue), Post-Wash (black), and Base2-Perturb2 (green) phases.
- **b.** Mean percentage of successful trials.
- **c.** Initial deviation in pointing angle during the Perturb1 (blue) and Perturb2 (green).
- **d.** $|\Delta PA|$ across epochs. Blue and green dots indicate trials after failure ($S = 0$) from the first and the second training sessions, respectively. Gray dots indicate trials after success ($S = 1$).
- **e.** Probability distribution of $|\Delta PA|$ (top panels) and $\Delta PA$ (bottom panels) from the first and the last epochs during the Perturb1 and Perturb2 phases. Values inside the bottom panels show the mean and SEM of $\Delta PA$ across participants.
- **f.** Relationship between $|\Delta PA|$ ($S = 0$, first epoch) and initial deviation for the Perturb1 (blue) and the Perturb2 (green).
- **g.** Relationship between the magnitude of changes in $|\Delta PA|$ ($S = 0$, first epoch) and changes in initial deviation from the first to the second training sessions. Dashed lines indicate regression line.
Fig 5. Increased exploration remained elevated during longer repetition of trials. a. Pointing angle while training the task. b. Mean percentage of successful trials. c. $|\Delta PA|$ after failure ($S = 0$, blue) and success ($S = 1$, gray) trials across epochs. Values show the mean and SEM for each trial or 40-trial epoch.

Fig 6. Non-learners showed no changes in exploratory behavior. a. Pointing angle during task training. Dots and shaded area show the mean and SEM for each trial during the Base1-Perturb1 (orange), Post (black) phases. Only the data collected in common across all the participants (i.e., Base1, Perturb1 and Post) are presented. b. Mean percentage of successful trials. d. $|\Delta PA|$ for each epoch. Orange and gray dots indicate trials after failure ($S = 0$) and success ($S = 1$), respectively. d. $|\Delta PA|$ between trials $n$ and $n + 1$ as a function of the feedback history for the three most recent trials [$S(n)$, $S(n - 1)$, $S(n - 2)$]. Orange and blue lines indicate data from the Non-learner and the Learner group, respectively.
<table>
<thead>
<tr>
<th></th>
<th>Block1</th>
<th>Block2</th>
<th>Block3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td>251.8 ± 63.3</td>
<td>198.9 ± 40.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(18.3 ± 8.5 %)</td>
<td>(12.5 ± 7.6 %)</td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td>222.9 ± 35.8</td>
<td>197.5 ± 28.8</td>
<td>191.6 ± 36.0</td>
</tr>
<tr>
<td></td>
<td>(13.1 ± 6.3 %)</td>
<td>(9.8 ± 6.3 %)</td>
<td>(11.2 ± 6.9 %)</td>
</tr>
<tr>
<td><strong>Experiment 3</strong></td>
<td>257.5 ± 52.6</td>
<td>214.9 ± 39.1</td>
<td>207.7 ± 51.0</td>
</tr>
<tr>
<td></td>
<td>(15.4 ± 4.3 %)</td>
<td>(9.2 ± 5.4 %)</td>
<td>(11.1 ± 6.5 %)</td>
</tr>
<tr>
<td><strong>Experiment 4</strong></td>
<td>224.2 ± 25.1</td>
<td>199.6 ± 31.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10.9 ± 3.3 %)</td>
<td>(7.6 ± 4.8 %)</td>
<td></td>
</tr>
<tr>
<td><strong>Non-learners</strong></td>
<td>264.4 ± 54.8</td>
<td>221.2 ± 61.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(18.8 ± 7.2 %)</td>
<td>(12.1 ± 7.4 %)</td>
<td></td>
</tr>
</tbody>
</table>

Values indicate the average (± SD) movement time in msecs (top) and percentage of number of trials outside the time window (bottom) in each block. Note that participants included in Experiment 1-3 are only those classified as learners.