eLife

Reviewed Preprint
v1 • March 18, 2026
Not revised

For correspondence:
adrianhaith@gmail.com

Competing interests: No competing interests declared

Reviewing editor: Jean-Jacques Orban de Xivry, KU Leuven, Belgium

Policy-Gradient Reinforcement Learning as a General Theory of Practice-Based Motor Skill Learning

Adrian M Haith
Johns Hopkins University, Baltimore, United States

eLife Assessment

This **valuable** computational study presents a conceptually simple and biologically plausible reinforcement-learning framework for motor learning based on policy-gradient methods. The evidence supporting the conclusions is **convincing**, including rigorous mathematical derivations of learning rules for the mean and variance of motor commands and simulation results for three sets of experimental data, based on three different motor learning tasks from the literature. However, there is a lack of a clear description of the specific conditions under which this framework yields unique mechanistic insights or predictive values, hence falling short of qualifying as a "general theory of motor learning". The work will be of interest to researchers in computational motor learning and motor neuroscience.

https://doi.org/10.7554/eLife.109934.1.sa2

Abstract

Mastering any new skill requires extensive practice, but the computational principles underlying this learning are not clearly understood. Existing theories of motor learning can explain short-term adaptation to perturbations, but offer little insight into the processes that drive gradual skill improvement through practice. Here, we propose that practice-based motor skill learning can be understood as a form of reinforcement learning (RL), specifically, policy-gradient RL, a simple, model-free method that is widely used in robotics and other continuous control settings. Here, we show that models based on policy-gradient learning rules capture key properties of human skill learning across a diverse range of learning tasks that have previously lacked any computational theory. We suggest that policy-gradient RL can provide a general theoretical framework and foundation for understanding how humans hone skills through practice.

Introduction

Learning a new motor skill invariably requires extensive, repetitive practice before we become proficient at it. Through practice, we gradually refine execution of motor skills through a combination of fine-tuning the exact motor commands we send to our muscles, and reducing the variability of these motor commands from movement to movement[1–3]. The learning principles underlying this gradual improvement are not clearly understood. Here, we show that practice-based motor learning across a variety of tasks is well captured by policy-gradient reinforcement learning, a simple and well-established model-free reinforcement learning (RL) method that is widely used in robotics and in many other continuous control settings[4–6].

Most existing computational theories of motor learning are based on principles of *supervised learning*. According to these theories, observed task errors (e.g. missing a target to the left/right) are used to guide compensatory changes in motor output[7–9]. Doing this successfully, however, depends critically on pre-existing knowledge about the structure of the task in order to translate observed task errors into appropriate changes in motor commands[10–12], but this knowledge may

not always be available when learning a new skill. These *error-based* learning models account well for how people rapidly adapt already familiar well-practiced movements to compensate for imposed perturbations[9,13,14], but cannot explain how brand new skills are learned and refined over hours of practice. Furthermore, error-based learning models account only for overall shifts in motor output and offer no account of the changes in variability that are a key characteristic of practice[2,15,16]. Alternative frameworks besides supervised learning are therefore needed to understand practice-based learning of motor skills.

Here, we show that motor skill learning through practice can be described using principles of *reinforcement learning* (RL), specifically, policy gradient methods. Policy gradient is a simple, model-free RL method that requires no prior task knowledge and guides improvements in performance based solely on observed states, actions, and a "reward" signal which provides a simple, scalar assessment of task success. We show how policy-gradient learning rules naturally account for changes in the mean and variance of behavior across a range of established skill-learning paradigms that have previously lacked any computational theory.

Results

Policy Gradient RL Framework

In general, RL describes the scenario in which an agent must learn from experience how to act in order to optimize an unknown reward function r. RL algorithms typically specify a *control policy* in the form of a probability distribution $\pi_\theta(a|s)$, over potential actions a, given current task state s.

This policy depends on adjustable parameters θ (which typically describe the mean and variance of selected actions) and the goal of RL is to determine policy parameters θ that maximize the overall expected reward $\mathbb{E}[r(s, a)]$. Pol*icy gradient* methods, attempt to optimize reward by following the gradient of the expected reward function with respect to the parameters $\theta, \frac{\partial}{\partial\theta}\mathbb{E}_{\pi_\theta}[r(s,a)]$. It is not possible to determine this gradient exactly, since the reward function is not known. However, the policy gradient theorem shows that this gradient equivalently be expressed as

$$\frac{\partial}{\partial\theta}\mathbb{E}_{\pi_\theta}[r(s,a)] = \mathbb{E}_{\pi_\theta}\left[r(s,a)\frac{\partial}{\partial\theta}\log\pi_\theta(a|s)\right], \quad (1)$$

which, since $\frac{\partial}{\partial\theta}\log\pi_\theta$ is known, can be straightforwardly approximated in a Monte Carlo manner by sampling actions and associated rewards. If this gradient estimate is based on just a single sample of actions and outcomes from one trial, this leads to the following update rule:

$$\theta_{i+1} = \theta_i + \alpha(r(a_i, s_i) - \bar{r}_i)\frac{\partial}{\partial\theta}\log\pi_\theta(a_i|s_i). \quad (3)$$

which describes how policy parameters θ on trial i should be updated based on the state s_i, action a_i, and reward outcome $r(a_i, s_i)$. Note that here, the raw reward term $r(a_i, s_i)$ in Equation (2) has been adjusted by $[r(a_i, s_i) - \bar{r}]$, where $\bar{r}$ is a running estimate of the recent rewards. This baseline subtraction is commonly employed in policy-gradient applications since it reduces the variance of the gradient estimate when using small numbers of trials.

This update rule, first introduced in 1992 under the moniker REINFORCE[5], serves as the foundation of policy-gradient methods for RL. The exact form of the update depends on the nature of the policy distribution $\pi_\theta(a_i|s_i)$ and its parameters θ. We consider policies that take the form of a multivariate Gaussian distribution over actions, with $\pi(a|s) \sim N(\mu, \Sigma)$. In this case, the policy-gradient update for the mean μ becomes:

$$\mu_{i+1} = \mu_i + \alpha(r_i - \bar{r}_i)\Sigma_i^{-1}(a_i - \mu_i), \quad (4)$$

where μ_i and Σ_i are the policy mean and covariance on trial i.

This simple learning rule is illustrated in Figure 1A, which depicts an arbitrary reward function on a 2-D action space. The update rule in Eq. (4) shifis the mean of the distribution towards the sampled action a_i, with a sign and amplitude that depends on the RPE, $(r_i - \bar{r}_i)$. If the reward

obtained is beGer than the average of recent previous attempts (i.e. if the RPE is positive), the policy mean μ will be updated towards action a_i (Fig. 1B). Conversely, if the outcome is worse than the average past behavior (negative RPE), the policy mean will become shifted away from action a_i (Fig. 1C).

In addition to updating the mean of the policy, Eq. (2) can also be applied to derive a learning rule for the covariance matrix of the policy, Σ. This can be represented as a diagonal matrix of log-eigenvalues v:

$$\Sigma = \begin{pmatrix} e^{v_1} & 0 & \cdots \\ 0 & e^{v_2} & \\ \vdots & & \ddots \end{pmatrix},$$

which ensures a valid covariance matrix for any value of (v_1, v_2). In this case, the general policy-gradient update for parameter v_k on timestep i becomes:

$$v_{k,i+1} = v_{k,i} + \frac{1}{2}\alpha(r_i - \bar{r}_i)\left[(a_{k,i} - \mu_{k,i})^2 - 1\right]. \qquad (5)$$

Parametrizing the covariance matrix in this way assumes that the actions along each dimension are sampled independently, but it is also possible to derived updates for learning covariance between actions.

Equation 3 above provides a general, trial-by-trial learning rule that can be applied to any task for which a parametrized, differentiable policy is specified. As a model of learning, its main free parameter is the learning rate α (though this can, in principle, take different values for different policy parameters). We show how this simple framework yields models of learning that align closely with human behavior across a range of very different tasks.

Application to a skittles-toppling throwing task

We first applied the policy-gradient learning rules described above to simulate learning in the skittles-toppling task developed by Müller and Sternad[17]. This well-studied task is inspired by the real-world bar game of skittles in which participants must throw a ball, suspended by a string from a supporting central post, towards a target skittle, aiming to topple it (Fig. 2A). In laboratory versions of this task, it has been found that people gradually learn to improve their performance over hundreds to thousands of trials of practice[17,18]. Participants' attempts on each trial are characterized by the launch angle and velocity of the ball, and only certain combinations of these actions will successfully topple the skittle (Figure 2B). Participants tend to initially attempt throws that follow a broad distribution in this action space but their actions become gradually refined with practice, first by shifting the mean of the distribution towards more successful actions, followed by more gradual reduction in the variability of actions and alignment with the shape of the reward landscape[2,17–19] (Figure 1B).

We modeled behavior in this task using the policy-gradient RL learning rule in Equation (2). As in the example in Figure 1, we represented the model as a multivariate Gaussian distribution with mean μ and covariance matrix Σ. In addition to representing the log-eigenvalues of the covariance matrix, we also introduced a further parameter that describes the covariation between actions through a rotation of the covariance matrix Σ. This resulted in five parameters in total: 2 for the mean and 3 for the covariance. We used a reward function equal to the minimum distance between the ball and the skittle, and used the same learning rate, $\alpha = 0.07$, for all parameters (chosen so that simulations approximately matched the timescale of human learning). Similar to human participants, the policy-gradient RL model exhibited an initial shift in mean towards more successful actions, followed by a gradual reduction in variance and alignment of the covariance matrix with successful regions of the action space (Fig. 2C).

To compare behavior between our model and human learners in more detail, we applied a well-established analysis[18,19] that characterizes the progression of different aspects of participants' behavior. This analysis decomposes an individual's performance in a given block into three distinct components by which it deviates from best-possible performance (Figure 2D): i) potential improvement obtained by translating the policy in action space (T-Cost), ii) potential

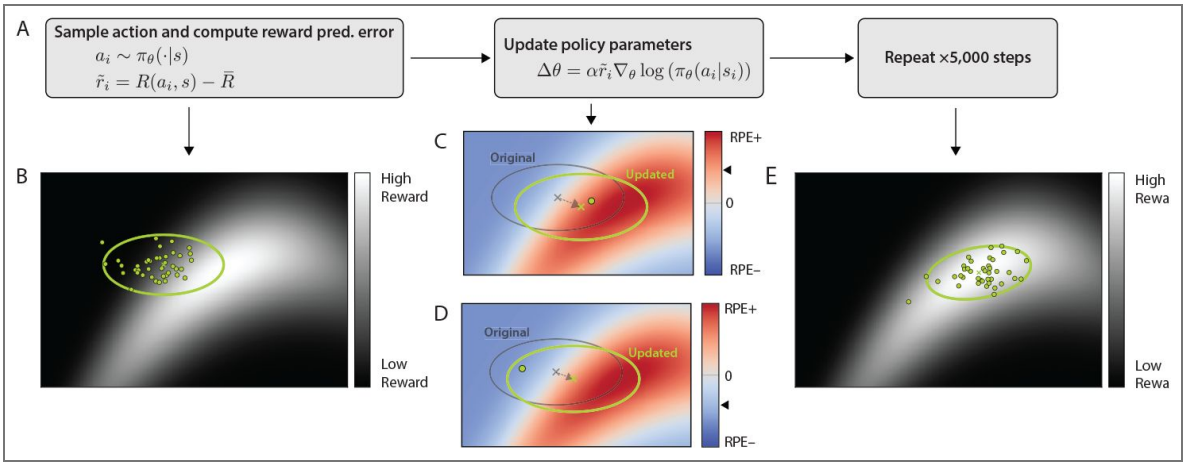

Figure 1. Illustration of policy-gradient RL.

A) Outline of policy gradient algorithm. B) Reward landscape for a task with 2-dimensional action space, and actions generated according to an initial policy, in this case a 2-dimensional Gaussian. The green ellipse indicates the covariance of the Gaussian policy (90% confidence interval), while dots represent individual samples. C) Illustration of policy update for a positive reward prediction error. In this trial, the randomly selected action leads to a beKer-than-average reward (positive reward prediction error). Consequently, the mean is updated towards the sampled action. D) Policy update for a negative reward prediction error. Here, the randomly selected action results in a worse-than-average reward (negative reward prediction error). Consequently, the mean is updated away from the sampled action. E) After a few thousand samples, the policy converges on the region of greatest reward.

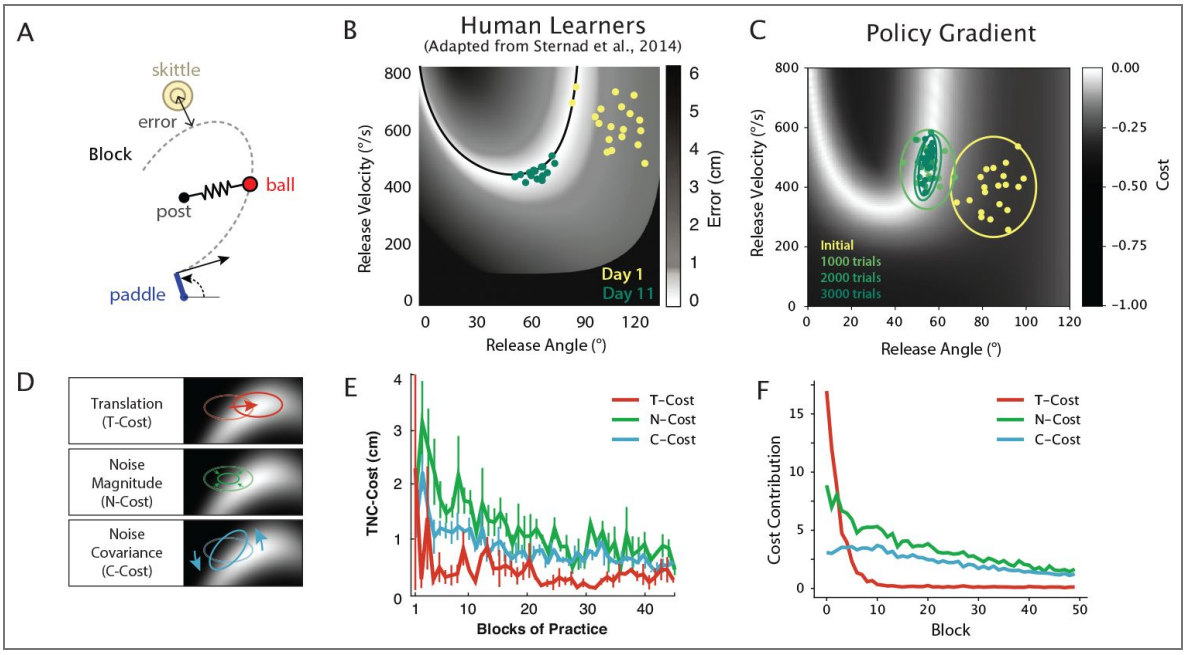

Figure 2. Policy-gradient RL model of a throwing task.

A) In Sternad and colleagues' virtual skiKles task[17,18], participants swivel a paddle and press a buKon to release a ball, aiming to topple a skiKle while avoiding a central post. The ball's movement is guided by an elastic force acting towards the central post, and is fully determined by the initial release angle and velocity. The task error is taken as the minimum distance of the balls trajectory to the center of the skiKle. B) Example human performance in this task: Heatmap shows task error given release angle and velocity, adapted from[18]. Yellow dots show example data from a human participant on their first day performing the task. Green dots show data from the same participant after 11 practice sessions (~2,000 trials). C) Example performance of the policy gradient algorithm trained on the same task. Ellipses show 90% confidence regions for the policy at different stages of learning. D) Illustration of the TNC-Cost decomposition used to compare human and model behavior. Based on actions taken over a block of 60 trials, the *Tolerance Cost* (T-Cost) estimates potential performance gains achievable over the current policy by translating the policy in action space. The *Noise-Amplitude Cost* (N-Cost) estimates potential performance gains achievable by uniformly scaling the noise. The *Noise-Covariance Cost* (C-Cost) estimates potential performance gains achievable by optimizing the policy covariance. E) TNC-Cost decomposition[19] for 9 participants, adapted from[18]. F) TNC-Cost decomposition for simulated learning of the policy-gradient RL model.

improvement obtained by scaling the overall variance of the policy (N-Cost), and iii) potential improvement obtained by altering the covariance between actions (C-Cost). This analysis has revealed that human learning proceeds by rapid initial translation of the mean action (reductions in the T-Cost), accompanied by more gradual shrinking of variance (reductions in the N-Cost) and alignment with reward landscape (C-Cost), with a slightly greater contribution from the N-Cost (Figure 2E). The policy-gradient RL algorithm exhibited almost identical characteristics to human behavior (Figure 2F). The only major difference was that the model exhibited slightly slower reduction of the T-cost in the model compared to human learners. This indicates slower convergence on a suitable mean action and is likely aGributable to a poorer initial policy, which was random for our model but in human learners could be informed by an initial intuitive understanding of the structure of the task. Nevertheless, the overall paGern of practice-based improvement was well-captured by policy-gradient RL.

In the skittles task, therefore, policy-gradient RL provided a very plausible and parsimonious model of how humans learn to improve their performance through practice, yielding successful learning over a similar timescale and with analogous characteristics to human learners.

Application to a "de novo" cursor control task

We next applied the policy gradient learning rule to a "*de novo*" cursor control learning task. In this task, participants have to maneuver a cursor using movements of the left and right hand through a non-intuitive mapping (Figure 3A). This task requires participants to learn a new controller "*de novo*", rather than by adapting an existing policy[20]. Participants initially find it extremely challenging to reach the targets with the cursor, but they gradually improve their performance over thousands of trials of practice across multiple sessions (Figure 3C).

To model behavior in this task using policy-gradient RL, we assumed that individuals generated point-to-point movements of their left and right hands, and systematically varied the direction and extent depending of these movements depending on the target location. The action space was therefore 4-dimensional. To allow for a policy where the actions could vary based on the target angle (state, s), the mean action was constructed as a weighted sum of local basis functions, $\mu(s) = ?_{-} w_{-}\Phi_{-}(s)$, yielding a mean action that depended smoothly on the state. We updated the weights w_{-} based on the learning rule derived from Equation (4). We allowed the estimate of recent reward, $\bar{r}$, driving the learning to also depend on the target direction (state s) using the sam e basis functions (making it a *value function* in RL terminology). In addition to learning suitable mean actions, we assumed distinct (state-independent) variances σ_k along each dim ension of the action space, and also learned these param eters via policy-gradient updates according to Equation (5). As a cost function, we used the distance between the end cursor location and the center of the target, and w e set the learning rate to $\alpha = 0.1$ in order to approximately match the timescale of human behavior.

Despite the added complexity of the policy being target-dependent and in a higher-dimensional, redundant action space, the policy-gradient model was able to learn the task within the same tim escale as human participants – around 2,500 trials[20] (Fig. 2B). Initial performance of the policy-gradient model, which was initialized with random weights, was poor, how ever. Performance, assessed in terms of absolute initial directional error, did not reach initial hu man performance levels until after around 1,000 trials of practice. After this point, how ever, learning of the model resembled human learning quite closely. Therefore, although initial performance of human learners was far better than the random initializations used in our policy-gradient RL model, likely due to people's explicit understanding of the structure of the task, the gradual, practice-based improvement over the next 2,000 trials may have occurred through policy-gradient RL.

To overcome the discrepancy in initial performance between the policy-gradient model and human learners, and thereby better compare the nature and time course of practice-based improvement, w e set the policy-gradient RL model's initial policy to match early policies of individual hu man participants, estimated based on their behavior in the first 120 trials. We then

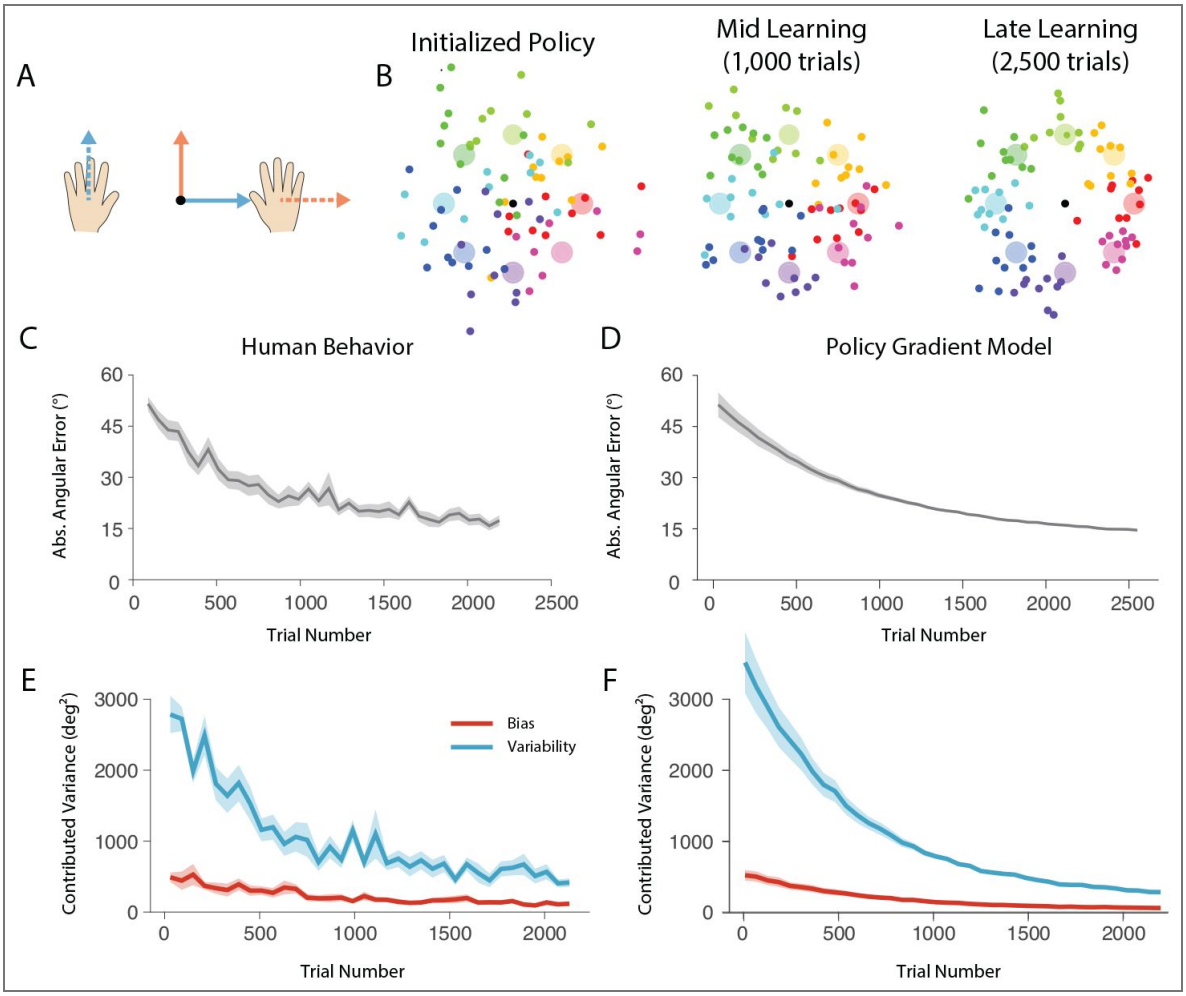

Figure 3. Policy-Gradient RL model of learning a cursor-control task.

A) In the bimanual cursor control task, participants must learn to control a cursor using a non-intuitive mapping from the positions of their left and right hands to an on-screen cursor. Forward-backward movement of the left hand leads to left-right movement of the cursor, and left-right movement of the right hand leads to forward-backward movement of the cursor. B) Learning of a policy-gradient reinforcement learning model for this task. Large circles represent different target locations, to be reached by the cursor from a central starting location. Each dot represents a trial endpoint, with color indicating the associated target for that trial. The three panels illustrate the policy at the outset of training, after 1,000 training trials, and after 2,500 training trials. C) Human performance in this task. Each curve shows average absolute directional error across blocks of 60 trials, averaged across N=13 participants. Shaded region indicates +/-standard error in the mean across participants. D) Average performance of the policy-gradient RL models with policies initialized to match initial behavior of human participants. E) Bias-variance decomposition for human performance over trials, averaged across participants. Most practice-based improvement is aKributable to reductions in variability. F) Averaged bias-variance decomposition for policy-gradient RL models with policies initialized to match initial behavior of human participants. Shaded regions indicate +/- sem across participants (E) or across models initialized to different participants (F).

sim ulated learning via policy-gradient given that subject-specific initialization. The learning simulated this way was closely aligned with human behavior, seen in the averaged time course over which the initial directional error improved with practice (Figure 3C,D).

We further analyzed human behavior in more detail by decomposing the total variance of participant's initial angular errors into a component due to bias of their movements (squared average angular deviation from the target) and a component due to variability around the mean deviation. Fig. 3E shows the contribution of these two components, revealing that the bulk of participants' improvement occurred through reduction of the variability of their movements. We applied the same analysis to learning simulated under the policy-gradient RL model and found that, with suitable parameter settings (α_{μ} = 0.1, $\alpha_{)}$ =0.02), the time course and nature of improvement were closely aligned with human behavior (Fig. 3F).

Therefore, similar to the skittles task, the practice-based component of learning in this task was well-captured by the principles of policy-gradient reinforcement learning.

Application to a precision motor execution task

In the first two tasks we considered, the learning challenge is clearly one of action *selection*, and these tasks are thus are naturally modeled within the RL framework which is fundamentally concerned with learning to *select* suitable actions. In contrast, many other practice-based motor learning tasks are thought to require learning to *execute* actions better, a capability referred to as "motor acuity", by analogy to visual acuity. One example of such a task is the "arc task"[16], in which participants use wrist movements to maneuver an on-screen cursor through an intuitive but unpracticed mapping, similar to using a laser pointer. Participants must guide the cursor through a narrow, semi-circular channel to reach a target circle (Fig. 4A). Initially, participants' trajectories are very variable, with a high proportion of failed trials in which the cursor leaves the channel. After around 1,000 trials of practice over multiple days, however, participants gradually become able to reduce the variability of their actions to improve the proportion of trials that remain inside the channel[16] (Figure 4A, right panel).

Because the average cursor path does not change much and only the variability across attempts changes substantially with practice, this task has been thought to purely assess how people learn to *execute* a movement better with practice, reducing their motor execution noise. However, since the timescale of learning (around 1,000 trials) is not too different from that in the skittles and cursor control tasks, we wondered whether learning in this task might in fact be better understood as one of action selection. Although the required path for the cursor appears obvious, the mapping between wrist movements and the cursor location was not very familiar to participants and they could not have known the exact movements required. We therefore developed a model of learning in this task to test whether practice-based improvements in performance could be accounted for by the same learning principles that could explain practice-based improvements in the skittles and cursor-control tasks.

To simulate execution of individual movement trajectories in this task, we adopted a model in which movements are generated based on a pre-planned sequence of subgoals which are updated every 130 ms. This previously proposed control scheme is consistent with observed micro-structure of behavior when generating complex trajectories[16,21]. Therefore, although the movements themselves are continuous, they can be conceived as being generated based on a discrete sequence of 2-D subgoals. We supposed that this sequence of subgoals may be the higher-level actions that must be selected and optimized in order to generate successful movements. Therefore, we developed a policy-gradient RL model in which the action a specified the pre-planned sequence of subgoals (six 2-d goal locations for a movement lasting 780 ms), giving a 12-dimensional action space. The policy took the form of a Gaussian distribution over this action space, with 12 parameters specifying the mean of the control points (six 2-D locations in space), and a further 12 parameters specifying their variance along each axis (log-eigenvalues of a diagonal covariance matrix). We used a reward function that combined a number of considerations related to task success: reaching the correct endpoint, remaining inside the

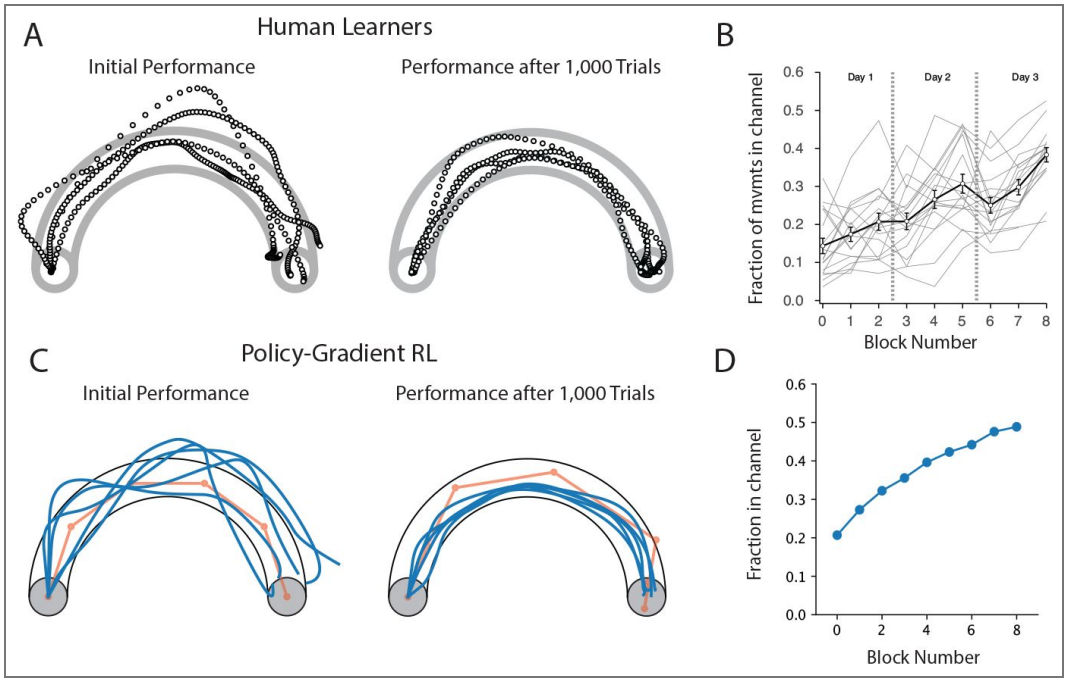

Figure 4. Policy-gradient model of learning to generate precise movements.

A) Precision execution task, adapted from[16]. Participants used movements of their wrist to rapidly guide a cursor through an arc-shaped path from left to right with the goal of not leaving the channel. Initially, participants' paths were highly variable and most trials were unsuccessful. After 8 sessions of practice (~1,000 trials), participants' movement variability was reduced, and the rate of successful movements increased. B) Fraction of trials successfully remaining within the channel throughout across 8 blocks of practice (120 trials per block), averaged across human participants. Panels A) and B), reproduced from[16]. C) Policy-gradient RL model of motor acuity learning. Initial policy was set to approximately match initial human performance. After 1,000 training trials, performance of the model was substantially improved. Red dots indicate intermediate goal locations used to model generation of the movement trajectories. The parameters of the policy were the 2-D location of the 6 goal locations, along with variance of their locations across trials. D) Fraction of trials staying within the channel, averaged across 100 simulated experiments.

channel, and maintaining a smooth trajectory. We used a fixed learning rate of $\alpha = 0.2$ for both mean and variance parameters, selected to approximately match the time course of human behavior.

Trajectories generated by the model looked qualitatively similar to those exhibited by human participants performing this task (Fig. 4C). Over 1,000 trials, the trajectories became progressively less variable until they most remained within the channel. We compared the time course of improvement over 8 blocks of 120 trials each between human participants (data from[16]), and our policy-gradient RL model, with the learning rate set to $\alpha = 0.2$. The model paralleled participants' steady improvement over the course of 8 blocks (960 trials).

Therefore, the policy-gradient RL framework not only successfully describes learning in tasks that obviously represent an action-selection challenge, the same principles can explain learning in tasks that appear to stress action execution, suggesting that apparent improvements in execution may in fact really be improvements in action selection driven by policy-gradient RL. This suggests that policy-gradient RL may be a very general principle underlying practice-based improvements in skilled motor performance.

Discussion

Despite many decades of effort developing computational models of human motor learning, we have no clear theory of how most everyday skills are learned. Here, we have shown that a single learning rule (Equation 2) based on policy-gradient RL accounts well for paGerns of practice-based improvement in a diverse range of motor skill learning paradigms. We therefore suggest that policy-gradient RL provides a plausible and promising general theory of practice-based motor skill learning.

Reinforcement learning has long been thought to be important for motor learning[22], but few concrete computational models have been proposed. In previous work, RL-inspired algorithms have been applied to explain human motor learning in highly simplified tasks. Several studies have examined learning in 1-dimensional learning tasks in which subjects must learn to generate a specific action (e.g. reaching in a specific, but unknown, direction) given success/failure feedback about their performance[23–27], and computational models have been proposed to account for behavior in these tasks. These models can mostly be construed as variants or approximations of policy-gradient learning rules[24–26]. Policy-gradient-like learning rules have also been successfully applied to explain locomotor adaptation[28].

An important difference between the models presented here and previous work on reinforcement-based motor learning is that previous experiments and theories have mostly emphasized binary success/failure feedback. Learning to adapt one's behavior under binary feedback is extremely challenging[29], and becomes particularly difficult in higher-dimensional tasks[30]. In the models presented here, learning is driven instead by a scalar reward function that takes continuous values reflecting the relative quality of movement. Importantly, this scalar reward function might not directly reflect overt rewards associated with success or failure, like whether the skittle was toppled or not, but reflect a more continuous grading of outcomes, like distance to the skittle. While overt rewards have been found to influence the rate of motor learning in certain tasks[31–33], these overt rewards appears to exert more of a modulatory effect, rather than serving as the main signal that drives learning.

One consistent finding in reward-driven motor learning tasks is that participants actively modulate their variability, increasing variability following errors[24,34,35], seemingly to promote exploration. This increase in variability is not predicted by the basic policy-gradient RL update in Equation 2. However, it is possible that this phenomenon is an artifact of using tasks in which learners must rely only on binary feedback. With binary rewards, outcomes from a single trial carry liGle information about the gradient of the overall reward function and increased variance can improve a learner's ability to reliably estimate the gradient of the reward function given few

actions. With continuous rewards, there is less benefit to increasing variability, since even small variability in actions can carry reliable information about the direction of the reward gradient, provided the relative differences in rewards for different actions can be reliably discriminated.

Motor learning is often considered to be a highly cognitive enterprise[36–39], with learning thought to involve a combination of reasoning about possible motor solutions to a task, and refinement of promising solutions[40]. The policy-gradient RL models presented here offer a concrete theory of the refinement part of this process. In this theory, however, the refinement occurs through a very simple learning rule that appears to have liGle need for cognitive involvement. The importance of cognitive contributions in the tasks we considered is clearly apparent in early learning in both the skittles task and the bimanual cursor control task, where initial human performance far surpassed that of randomized initial policies used in our models. Policy-gradient RL models initialized with random policies could require hundreds of trials of training before catching up to initial human performance levels. After that point, however, subsequent learning of the model was closely aligned to human learning. The superior initial performance of human learners compared to the model was likely aGributable to high-level understanding of the structure of the task and the ability to reason about potential solutions that would fare much better than picking an initial policy at random. The fact that the gap between human learners, and our simple policy-gradient RL models was only really apparent at the outset of learning suggests that cognitive contributions to motor skill learning may be restricted only to the very early phases of learning, and there may be liGle need for cognitive involvement for subsequent, practice-based improvement.

The limited role for cognition in practice-based learning is in contrast to many theories of motor learning, which posit that cognition serves to guide action selection throughout the course of learning[37]. It also appears to contradict extensive literature in skill acquisition which has argued that gaining expertise in complex skills requires *deliberate prac7ce* – a concerted and cognitively demanding approach to practice[41,42]. Can the notion of cognitively-demanding deliberate practice be reconciled with a policy-gradient RL theory of practice? One possible role for cognition in practice, within the framework of policy-gradient RL, is that it may help to shape the reward function. The exact nature of the reward function is a subjective aspect of our model, though in additional simulation (not shown), we found that basic properties of learning were preserved even if the exact form of the reward function was varied (e.g. using squared distance to the goal, rather than absolute distance). It is well established that primary rewards may not be the most efficient way to train an RL agent, and that alternative pseudo-reward functions constructed internally can be significantly more effective[43,44]. Intuitively, a 'good' outcome of a golf swing at the driving range would look very different for a novice versus an accomplished golfer. The reward function therefore ought to be adjusted throughout the course of learning in order to maximize learning efficiency, in contrast to the static reward functions used in our models. This progressive reshaping of the reward function could be achieved through explicit, cognitive processes that oversee learning and assume the role of 'critic' (in the sense of actor-critic RL), providing the reward signal driving learning in Equation 2. This 'cognitive critic' proposal is consistent with the emphasis in deliberate practice on ensuring high-quality feedback on the quality of one's performance and enhancing self-monitoring and assessment.

More broadly, the often slow nature of simple RL algorithms has led to the idea of meta-reinforcement learning[45,46] – the problem of how to optimize learning speed of RL agents by constructing a suitable curriculum, potentially including adjustments to the reward function as well as selection of specific states and tasks to focus practice on. Cognition in skill learning, then, may more generally be required to accelerate the slow nature of RL, without actually being responsible for reasoning about action selection other than at the very outset. These influences of cognition may also account for the often substantial individual differences in learning across individuals in skill learning, rather than differences in learning rate parameters.

Having a concrete, computational theory of motor skill learning will be beneficial for gaining greater clarity in identifying its neural basis. The policy-gradient model in general requires two distinct sets of computations, one to compute the reward prediction error (or, more generally, the value function or advantage function) and one to update the policy. It is well known that

dopaminergic neurons seem to compute a reward-prediction error, and are thus a clear candidate for computing the advantage function. Reward-related signals are observed throughout the brain[47]. Plasticity in the cortex is well known to be modulated by dopamine and dopaminergic projections to motor cortex are necessary for learning dexterous skills in rats[48–50]. Therefore, the policy itself is likely represented in motor cortical regions.

In conclusion, our results establish a concrete computational model of practice-based skill acquisition in humans that can describe learning across a diverse range of learning tasks that emphasize the importance of repeated practice. Given that policy-gradient methods are at the heart of recent advances in robotics[4,51], and have also been successfully applied to musculoskeletal control problems[52], it seems promising that policy-gradient RL could prove to be a universal principle that can account for the seemingly limitless range of skills that humans are capable of learning.

Methods

General Approach

For simplicity, we consider only single-timestep tasks in which an action or actions are selected only once, at the start of the movement, and these actions directly give rise to a reward. (We do not consider temporally-extended tasks in which the later states depend on prior actions, though the general policy-gradient framework is very much extendable to that setting).

We assume that users select actions a stochastically, based on the current state s, according to a policy specified as a probability distribution over actions, $\pi_\theta(a|s)$. Here, θ denotes parameters of the policy, which are adjusted through learning. For a given task, actions lead to an outcome which is then associated with a reward function $r(s, a)$, providing a scalar metric of performance. The goal of reinforcement learning is to identify policy parameters θ^* that maximize the average reward.

Policy gradient methods aim to optimize performance by adjusting the parameters θ by gradient descent. That is, they seek to follow the gradient,

$$\Delta\theta = \frac{\partial}{\partial\theta}\mathbb{E}_{\pi_\theta}[r(s,a)].$$

It is not immediately obvious how to calculate this expected gradient, since neither the expected reward nor its gradient with respect to θ is known. However the problem is simplified by the *policy gradient theorem*, which states that the gradient can equivalently be expressed as:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{\pi_\theta}[r(s,a)] = \mathbb{E}_{\pi_\theta}\left[r(s,a)\frac{\partial}{\partial\theta}\log\pi(a|s)\right].$$

Critically, this expectation can now be approximated by sampling a large number of actions and associated costs, giving rise to a batch update rule as is typically used in reinforcement learning applications. It is possible, however, to use a batch size of 1, which yields an online learning rule which we adopt as our model of trial-by-trial skill learning. Therefore, on trial i, we assume that the policy parameters θ are updated according to

$$\theta_{i+1} = \theta_i + \alpha\, r(s_i, a_i)\, \frac{\partial}{\partial\theta}\log\pi(a_i|s_i),$$

where α is a learning rate.

This learning rule can be adjusted to improve stability by replacing the reward in Eq.4.1 with a reward prediction error, i.e. subtracting out a baseline expected reward based on recent attempts:

$$\theta_{i+1} = \theta_i + \alpha(r(s_i, a_i) - \bar{r}_i).\frac{\partial}{\partial\theta}\log\pi(a_i|s_i).$$

In our simulations, we update the average reward, $\bar{r}_i$, trial-by-trial according to with β = 0.95.

$$\bar{r}_{i+1} = (1-\beta)\bar{r}_i + \beta r(s_i, a_i),$$

In all of our models, we assume that an individual's policy $\pi(s|a)$ follows a Gaussian distribution parametrized by a mean μ (which, may depend on the state s) and a covariance matrix Σ:

$$\pi(a|s) = 2\pi^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}e^{\frac{1}{2}(\mu-a)^{\top}\Sigma^{-1}(\mu-a)},$$

where N is the dimensionality of the action space. Updates to the mean in this case are given by:

$$\mu_{i+1} = \mu_i + \alpha\Sigma^{-1}(\hat{a}_i - \mu_i)\,\tilde{r}(a_i).$$

Policy-gradient methods are known to be vulnerable to instability triggered by occasionally large parameter updates. This issue is exacerbated when updates are based on single action-outcome observations, rather than batches, as is typical. A widely used method to avoid such instabilities is a method called proximal policy optimization[51], which limits the size of parameter updates in a way that bound the possible ratio of probabilities of sampled actions between the original and updated policies. In line with this approach, we bounded updates to the mean to ensure that:

$$|\Delta\mu| \leq \frac{\epsilon}{|\Sigma^{-1}(a_i-\mu_i)|},$$

where ϵ is a small number, usually around 0.1. This corresponds to a linear approximation of the PPO bound for the mean of a Gaussian policy.

Skittles task model

In the virtual skittles task, participants must launch a virtual ball towards a target skittle, while avoiding a central obstacle. The ball's dynamics are governed by a springlike aGractor towards the central obstacle. Participants' actions in a given trial are the launch angle and velocity with which the ball is released, i.e., $a = [v, \Phi]$[6]. The skittles task was modeled as described in[18] and as illustrated in Fig. 2A. We assumed a paddle of length $\boldsymbol{l}$=0.4 m, situated 1.3 m below the central post, and with the skittle located 0.25 m to the right and 1.2 above the central post. We assumed a spring-like force aGracting the ball towards the center of the task space, with stiffness $\boldsymbol{k}$ = **1** N/m, and with the ball's mass equal to **0. 1** kg. To calculate the ball's trajectory, we determined the initial release position and velocity of the ball:

$$x_0 = \begin{pmatrix} l\cos\phi_i \\ l\sin\phi_i \end{pmatrix}, \quad \dot{x}_0 = \begin{pmatrix} v_i\sin\phi_i \\ v_i\cos\phi_i \end{pmatrix}.$$

The full trajectory of the ball was then given by:

$$x_t = x_0\cos\omega t + \frac{v_0}{\omega}\sin\omega t$$

where $\omega = \sqrt{k/m}$. We determined the cost on a given trial by simulating the trajectory of the ball, and calculating the minimum distance between the ball and the center of the skittle, which served as the reward signal in our model:

$$R(v,\phi) = -\min_t(|\text{ball}(t) - \text{skittle}(t)|).$$

The learner's policy was represented by a mean action μ = (v, Φ)[6] and a covariance matrix Σ. To ensure that rotations of the covariance were meaningful despite the two different action dimensions being qualitatively different and potentially measured in different units, we represented the policy using a normalized, unitless co-ordinate system by subtracting the initial mean and dividing by the initial standard deviation for each action dimension:

$$\hat{a} = \begin{pmatrix} (v-v_0)/\sigma_0^v \\ (\phi-\phi_0)/\sigma_0^\phi \end{pmatrix},$$

where v_8 is the initial mean release velocity, Φ_8 is the initial mean release angle, and σ_0^v and σ_0^ϕ: are the initial standard deviations. The initial normalized policy mean was then simply μ_8 = (0,0)[6] and with initial covariance matrix was given by the identity matrix, i.e. $\Sigma_8 = \boldsymbol{I}$.

We parametrized Σ using an eigenvalue decomposition,

$\Sigma = Q_\phi \Lambda Q_\phi^\top$

where Q: is a rotation matrix parametrized by angle Φ:

$$Q_\phi = \begin{pmatrix} \sin\phi & \cos\phi \\ -\cos\phi & \sin\phi \end{pmatrix}.$$

and Λ is a diagonal matrix of eigenvalues. To ensure numerical stability and avoid updates that could lead to negative eigenvalues (and an invalid covariance matrix), we represented this matrix in terms of its log-eigenvalues:

$$\Lambda = \begin{pmatrix} e^{\nu_1} & 0 \\ 0 & e^{\nu_2} \end{pmatrix}.$$

The overall policy was then specified by

$$\pi(\hat{a}) = \frac{1}{2\pi}|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{a}-\mu)^\top \Sigma^{-1}(\hat{a}-\mu)},$$

with five variable parameters in total, two for the mean Lμ_9, $\mu_<$M, and three for the covariance matrix (ν_1, ν_*, Φ).

The policy gradient update is determined by the gradient of the log-probability. For the mean, this is given by

$$\frac{\partial}{\partial \mu}\log\pi(\hat{a}) = \Sigma^{-1}(\hat{a} - \mu).$$

This leads to the following learning rule for updating the normalized mean action:

$$\mu_{i+1} = \mu_i + \alpha\Sigma^{-1}(\hat{a}_i - \mu_i)\,\tilde{r}(a_i).$$

Here, $\tilde{r}(a_i) = r(a_i) - \bar{r}_i$ is the reward prediction error, which depends on the average recent reward $\bar{r}_i$, which is updated after every trial according to:

$$\bar{r}_i = \beta\,\bar{r}_{i-1} + (1-\beta)r(a_i).$$

For the log-eigenvalues of the covariance matrix, the gradient of the log-probability is given by

$$\frac{\partial}{\partial \nu_k}\log\pi(a) = -\frac{1}{2} - \frac{1}{2}z_k^2 e^{-\nu_k},$$

where $z = Q_\phi^\top(\hat{a}_i - \mu_i)$ and z_+ is its k$^{\text{th}}$ element. This leads to the following learning rule for the log-eigenvalues ν:

$$\nu_{k,i+1} = \nu_{k,i} + \alpha\frac{1}{2}\left(z_k^2 e^{-\nu_{k,i}} - 1\right)[r(a_i) - \bar{r}_i].$$

Lastly, for the orientation of the covariance matrix (given by the angle variable Φ), the gradient of the log-probability is given by

$$\frac{\partial}{\partial \phi}\log\pi(a) = \mathrm{trace}\left((-Q^\top z z^\top \Lambda^{-1})^\top \frac{\partial Q}{\partial \phi}\right).$$

This leads to the learning rule:

$$\phi_{i+1} = \phi_i + \alpha\left(-Q^\top z_i z_i^\top \Lambda^{-1}\right)\begin{pmatrix} -\sin\phi_i & \cos\phi_i \\ -\cos\phi_i & -\sin\phi_i \end{pmatrix}\tilde{r}(a_i).$$

In all the above updates, α is a fixed learning rule shared across all parameters. Gradient calculations were confirmed using an automatic differentiation tool (pytorch).

For the TNC-Cost analysis, we grouped trials into blocks of 60 trials (to match human experimental data). For each block of 60 trials, we calculated the T-Cost, N-Cost and C-Cost as follows. For the T-Cost, we used a continuous optimization method to determine the best-possible reward that could be applied to each block of actions by applying a common translation in action space to all actions (Fig. 2D, top panel). The T-Cost was quantified as the difference between this optimized cost and the original cost. For the N-Cost, we used continuous optimization to determine the best-possible

reward obtainable by scaling the distance of each action from the mean across all actions. The N-Cost was quantified as the difference between this optimized cost and the original cost. For the C-Cost, we shuffled the pairings between different selected release velocities and angles to find the pairing that yielded the best-possible reward, which we identified through a greedy pair-swapping approach. The C-Cost was quantified as the difference between this optimized cost and the original cost.

De Novo Learning Model

The model for the de novo learning task follows very similarly to the model for the skittles task. In this task, however, we aim to learn a policy that varies based on the target location (represented by the state *s*). We assume that targets appear at a fixed distance *d* but at a variable angle *s* from the start location. Participants must learn appropriate displacements of the left and right hands a_L and a_R in order to bring the cursor to the target, so that the action *a* in each trial is a 2-d vector: $a = (a_L, a_R)$. We assumed that the final location of the cursor was given by $c = (a_L, a_R)$. $c = (a_L, a_R)$. The reward function was set to be equal to the negative of the absolute distance between the cursor and target location, i.e.

$$r(s,a) = -\sqrt{(a_L - d\cos(s))^2 - (a_R - d\sin(s))^2}.$$

The movement policy specifies a distribution of displacements for the left and right hands that is now dependent on target angle *s*. We again assume that this distribution is Gaussian and that the m ean displacements of the left and right hands are represented as a sum of Gaussian basis functions, i.e.

$$\mu_L = \sum w_j^L \phi_j(s); \mu_R = \sum w_j^R \phi_j(s).$$

Since the target could appear in any direction, we used a circular basis function taking the shape of the von Mises distribution, taking the form

$$\phi_i(s) = \frac{e^{-\kappa\cos(s-c_i)} - e^{-\kappa}}{2\pi I_0(\kappa)},$$

where κ is a concentration parameter (analogous to the reciprocal of the variance of a Gaussian) and $\boldsymbol{I}_8$ is the modified Bessel function of the first kind of order zero. We included 31 basis functions with centers c_i equally spaced around the circle, and set κ = 5, corresponding to a basis function width of approximately 25°. In this case, the policy mean is parametrized by w^{ti} and $w^{?}$. The policy gradient update for these parameters is given by:

$$w_{i+1,j}^L = w_{i,j}^L + \alpha_w \left(a^L - w_i^{L\top}\phi(s_i)\right)\phi_j(s_i)[r(a_i,s_i) - \bar{r}_i(s_i)].$$

For the covariance, we assume a single parameter, corresponding to the log-variance, that applies equally to left and right hand actions, and is independent of target direction, i.e.

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}.$$

Finally, since performance could be drastically different depending on the state, we maintained a state-dependent average reward (i.e. a value function) for each possible target state *s*. The update rule for σ is then given by

$$\sigma_{i+1} = \sigma_i + \alpha_\sigma \frac{z_i^2}{\sigma^2}\left[r(s_i,a_i) - \bar{r}(s_i)\right].$$

Policy-gradient RL models initialized with randomized weights had an initial directional much higher than human learners and initial learning was quite slow. To better mimic human performance, we initialized the model by matching the initial policy to human participants' early performance (blocks 2-3 of 35). In human experiments, participants were able to make online corrections and each trial only ended when the cursor reached the target. We couldn't base the initialized policy on participants' endpoint data. Instead, we constructed projected endpoints based on initial velocities of the left and right hands (taken 100ms after movement onset). We rescaled these initial velocities to convert them into projected endpoints, scaling all trials by a

fixed amount (per participant) determined such that the mean cursor displacement matched the distance to the target (12cm). We used least-squares regression with L_2 regularization (ridge regression, $\alpha = 1$) to determine initial policy weights. The residual variance around this learned policy was used to initialize v_i for each action dimension. The policy parameters were normalized based on the mean initial standard deviation for the two task-relevant action dimensions. Lastly, we used a separate, slower learning rate for the variance parameters, $\alpha_{FG} = 0.02$ to better align with human performance, which showed slower improved compared to when the policy-gradient variance learning rate was matched to that of the mean ($\alpha = 0.1$).

Arc task model

In the arc task, participants make rapid movements to guide a cursor through a narrow arc from a start location to an end location[16]. A feature of behavior in such tasks is the presence of highly regular sub-movements occurring with a highly regular period of 130 ms. Following a model proposed by Guigon[21], we model this behavior using an optimal feedback control model with intermiGently-updated movement goals. We model action selection in this task as selecting an appropriate sequence of abstract movement goals to achieve task success.

We model movements with duration 780 ms, so as to consist of 6 sub-movements of 130ms duration each. During each sub-movement, we assumed that movements pursued a fixed spatial goal $\mathbf{g}^i$ following an optimal controller with a horizon of 280 ms (a value which has been shown to account for the sub-movement structure of human movements). We modeled the wrist as a dynamical system with state $\mathbf{x}$ comprising position x, velocity x?, and two muscle states m_1 and m_2:

$$\mathrm{x}_t = \begin{pmatrix} x_t \\ \dot{x}_t \\ m_{1,t} \\ m_{2,t} \\ g \end{pmatrix}.$$

Here, g is the current goal location, which is also incorporated into the state, which allowing the endpoint cost to be expressed as $J = \mathbf{x}^6 Q\,\mathbf{x}$, with

$$Q = \begin{pmatrix} w_p & 0 & 0 & 0 & -w_p \\ 0 & w_v & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -w_p & 0 & 0 & 0 & w_p \end{pmatrix},$$

where w_H and w_9 are cost weights penalizing the squared deviation from the target, and squared velocity at the endpoint of the movement. The control cost is given by $J_{\mathbf{u}} = \mathbf{u}^6\mathbf{u}$.

The dynamics are given by

$$\dot{\mathbf{x}}_t = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ - & -\frac{l}{I} & \frac{1}{I} & 0 & 0 \\ 0 & 0 & -\frac{1}{\tau} & \frac{1}{\tau} & 0 \\ 0 & 0 & 0 & -\frac{1}{\tau} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{x}_t + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{\tau} \\ 0 \end{pmatrix} \mathbf{u}_t,$$

expressed more compactly as $\dot{\mathbf{x}}_t = A\mathbf{x}_t + B\mathbf{u}_t$ the above equation, $\boldsymbol{I}$ is the inertia of the wrist, l is viscosity, and τ is a time constant related to muscle activation.

We extend this 1-d model to two dimensions by doubling the number of states, and changing the matrices to be block-diagonal, i.e.

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_x \\ \mathbf{x}_y \end{pmatrix},\ A = \begin{pmatrix} A_x & 0 \\ 0 & A_y \end{pmatrix},\ B = \begin{pmatrix} B_x & 0 \\ 0 & B_y \end{pmatrix},\ Q = \begin{pmatrix} Q_x & 0 \\ 0 & Q_y \end{pmatrix}.$$

We discretized this problem using matrix exponentials, using a discretization timestep of 0.001s, and solved for the optimal time-varying control gains L_7 using the RicaN equations (see Todorov), so that the motor commands on timestep i were given by $\mathbf{u}_i = -L_i\mathbf{x}_i\,\mathbf{u}_i$.

To simulate a movement for a given sequence of N goals

$$g = (g_x^1, g_x^2, \ldots, g_x^N,\ g_y^1, g_y^2, \ldots, g_y^N),$$

we set the goal states in according to the initial goal location, and simulated the first 130ms of a 280ms movement using the optimal control gains L_i. After each 130ms, the goal states were updated and a new movement was simulated starting from the end state of the prior movement, and resetting the control gains to $L_1$for the new submovement.

For the policy gradient optimization, we took participants' action in each trial to be the specifaction of the sequence of goals **g** driving the movement. We assume that participants' policy was given by a multivariate Gaussian distribution:

$$\pi(g) \sim N(\mu, \Sigma).$$

We initialized μ as a sequence of equally-spaced points along the center of the arc. We initialized Σ as a diagonal matrix with equal variance for all dimensions, i.e. $\Sigma = \sigma_8 \boldsymbol{I}$.

We assumed that the reward function for successful movements comprised a number of components. First, we penalized deviations from the center of the arc:

$$r_{\text{arc}} = -\sum_i (\rho_i - c_i),$$

Where ρ_i reflects the radial position of the cursor on timstep *i*. We also included a term to encourage smooth trajectories by penalizing movement jerk:

$$r_{\text{jerk}} = -\sum_i (\ddot{x}_i^2 - \ddot{y}_i^2).$$

We also included a state reward at the end of the movement equivalent to the negative of the costs used in the optimal control problem $r_{\text{end}} = -\mathbf{x}_N^{\top} Q \mathbf{x}_N$ he total reward was then given by:

Policy updates were as per equations 4 and 5.

$$r = w_{\text{arc}} r_{\text{arc}} + w_{\text{jerk}} r_{\text{jerk}} + r_{\text{end}}.$$

All simulations were developed in python. Code for generating all simulations is available on the author's GitHub page: *https://github.com/adrianhaith/PolicyGradientSkillLearning*.

Data availability

Code for all simulations in the manuscript, along with all analyzed datasets, are available at: *https://github.com/adrianhaith/PolicyGradientSkillLearning*.

Acknowledgements

Thanks to Eric Grießbach, John Krakauer and Joshua Cashaback for helpful comments on the paper.

Additional information

Author ORCID iDs

Adrian M Haith: https://orcid.org/0000-0002-5658-8654

References

1. **Diedrichsen J.,** Kornysheva K. (2015) Motor skill learning between selection and execution. *Trends Cogn. Sci* **19**:227-233
2. **Sternad D.** (2018) It's not (only) the mean that maGers: variability, noise and exploration in skill learning. *Curr. Opin. Behav. Sci* **20**:183-195
3. **Haith A. M.,** Krakauer J. W. (2018) The multiple effects of practice: Skill, habit and reduced cognitive load. *Curr. Opin. Behav. Sci*

4. **Peters J.,** Schaal S. (2006) Policy Gradient Methods for Robotics. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2219-2225 https://doi.org/10.1109/IROS.2006.282564
5. **Williams R. J.** (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach Learn* **8**:229-256
6. **SuGon R. S.,** McAllester D., Singh S., Mansour Y. (1999) Policy Gradient Methods for Reinforcement Learning with Function Approximation.
7. **Berniker M.,** Kording K. (2008) Estimating the sources of motor errors for adaptation and generalization. *Nat. Neurosci* **11**:1454-1461
8. **Donchin O.,** Francis J. T., Shadmehr R. (2003) Quantifying Generalization from Trial-by-Trial Behavior of Adaptive Systems that Learn with Basis Functions: Theory and Experiments in Human Motor Control. *J. Neurosci* **23**:9032-9045
9. **Haith A. M.,** Krakauer J. W. (2013) Model-Based and Model-Free Mechanisms of Human Motor Learning. In: Richardson M, Riley M, Shockley K (Eds). *Progress in Motor Control* New York, NY: Springer. pp. 1-21 https://doi.org/10.1007/978-1-4614-5465-6_1
10. **Jordan M. I.,** Rumelhart D. E. (1992) Forward models: Supervised learning with a distal teacher. *Cogn. Sci* **16**:307-354
11. **Abdelghani M. N.,** Lillicrap T. P., Tweed D. B. (2008) Sensitivity Derivatives for Flexible Sensorimotor Learning. *Neural Comput* **20**:2085-2111
12. **Hadjiosif A. M.,** Krakauer J. W., Haith A. M. (2021) Did We Get Sensorimotor Adaptation Wrong? Implicit Adaptation as Direct Policy Updating Rather than Forward-Model-Based Learning. *J. Neurosci* **41**:2747-2761
13. **Cheng S.,** Sabes P. N. (2006) Modeling Sensorimotor Learning with Linear Dynamical Systems. *Neural Comput* **18**:760-793
14. **Shadmehr R.,** Smith M. A., Krakauer J. W. (2010) Error Correction, Sensory Prediction, and Adaptation in Motor Control. *Annu. Rev. Neurosci* **33**:89-108
15. **Georgopoulos A. P.,** Kalaska J. F., Massey J. T. (1981) Spatial trajectories and reaction times of aimed movements: effects of practice, uncertainty, and change in target location. *J. Neurophysiol* **46**:725-743
16. **Shmuelof L.,** Krakauer J. W., Mazzoni P. (2012) How is a motor skill learned? Change and invariance at the levels of task success and trajectory control. *J. Neurophysiol* **108**:578-594
17. **Muller H.,** Sternad D. (2004) Decomposition of variability in the execution of goal-oriented tasks: three components of skill improvement. *J. Exp. Psychol. Hum. Percept. Perform* **30**:212-233
18. **Sternad D.,** Huber M. E., Kuznetsov N. (2014) Acquisition of Novel and Complex Motor Skills: Stable Solutions Where Intrinsic Noise MaGers Less. In: Levin M. F. (Ed). *Progress in Motor Control* New York, NY: Springer. pp. 101-124 https://doi.org/10.1007/978-1-4939-1338-1_8
19. **Cohen R. G.,** Sternad D. (2009) Variability in motor learning: relocating, channeling and reducing noise. *Exp. Brain Res* **193**:69-83
20. **Haith A. M.,** Yang C. S., Pakpoor J., Kita K. (2022) De novo motor learning of a bimanual control task over multiple days of practice. *J. Neurophysiol* **128**:982-993
21. **Guigon E.** (2023) A computational theory for the production of limb movements. *Psychol. Rev* **130**:23-51
22. **Chen X.,** Holland P., Galea J. M. (2018) The effects of reward and punishment on motor skill learning. *Curr. Opin. Behav. Sci* **20**:83-88
23. **Wu H. G.,** Miyamoto Y. R., Castro L. N. G., Ölveczky B. P., Smith M. A. (2014) Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nat. Neurosci* **17**:312-321
24. **Dhawale A. K.,** Miyamoto Y. R., Smith M. A., Ölveczky B. P. (2019) Adaptive Regulation of Motor Variability. *Curr. Biol* **29**:3551-3562.e7

25. **Cashaback J. G. A.,** McGregor H. R., Mohatarem A., Gribble P. L. (2017) Dissociating error-based and reinforcement-based loss functions during sensorimotor learning. *PLOS Comput. Biol* **13**:e1005623
26. **Therrien A. S.,** Wolpert D. M., Bastian A. J. (2018) Increasing Motor Noise Impairs Reinforcement Learning in Healthy Individuals. *eNeuro* **5** https://doi.org/10.1523/ENEURO.0050-18.2018
27. **Izawa J.,** Shadmehr R. (2011) Learning from Sensory and Reward Prediction Errors during Motor Adaptation. *PLOS Comput. Biol* **7**:e1002012
28. **Seethapathi N.,** Clark B. C., Srinivasan M. (2024) Exploration-based learning of a stabilizing controller predicts locomotor adaptation. *Nat. Commun* **15**:9498
29. **Darshan R.,** Leblois A., Hansel D. (2014) Interference and Shaping in Sensorimotor Adaptations with Rewards. *PLoS Comput. Biol* **10**
30. **van der Kooij K.,** van Mastrigt N. M., Crowe E. M., Smeets J. B. J. (2021) Learning a reach trajectory based on binary reward feedback. *Sci. Rep* **11**:2667
31. **Vassiliadis P.,** et al. (2021) Reward boosts reinforcement-based motor learning. *iScience* **24**
32. **Abe M.,** et al. (2011) Reward improves long-term retention of a motor memory through induction of offline memory gains. *Curr. Biol. CB* **21**:557-562
33. **Galea J. M.,** Mallia E., Rothwell J., Diedrichsen J. (2015) The dissociable effects of punishment and reward on motor learning. *Nat Neurosci*
34. **Roth A. M.,** et al. (2023) Reinforcement-based processes actively regulate motor exploration along redundant solution manifolds. *Proc. R. Soc. B Biol. Sci* **290**:20231475
35. **Pekny S. E.,** Izawa J., Shadmehr R. (2015) Reward-Dependent Modulation of Movement Variability. *J. Neurosci* **35**:4015-4024
36. **Krakauer J. W.,** Hadjiosif A. M., Xu J., Wong A. L., Haith A. M. (2019) Motor Learning. *Comprehensive Physiology* 613-663 https://doi.org/10.1002/cphy.c170043
37. **Taylor J. A.,** Ivry R. B. (2012) The role of strategies in motor learning. *Ann. N. Y. Acad. Sci* **1251**:1-12
38. **Lee T. D.,** Swinnen S. P., Serrien D. J. (1994) Cognitive Effort and Motor Learning. *Quest* **46**:328-344
39. **Du Y.,** Krakauer J. W., Haith A. M. (2022) The relationship between habits and motor skills in humans. *Trends Cogn. Sci* **26**:371-387
40. **Tsay J. S.,** et al. (2024) Fundamental processes in sensorimotor learning: Reasoning, Refinement, and Retrieval. *PsyArXiv* https://doi.org/10.31234/osf.io/x4652
41. **Ericsson K.,** Krampe R., Tesch-Roemer C. (1993) The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev* **100**:363-406
42. **Ward P.,** Hodges N. J., Starkes J. L., Williams M. A. (2007) The road to excellence: deliberate practice and the development of expertise. *High Abil Stud* **18**:119-153
43. **Singh S.,** Lewis R. L., Barto A. G. (2009) Where Do Rewards Come From?.
44. **Weber L. A.,** Yee D. M., Small D. M., Petzschner F. H. (2025) The interoceptive origin of reinforcement learning. *Trends Cogn. Sci* **29**:840-854
45. **Narvekar S.,** Stone P. (2018) Learning Curriculum Policies for Reinforcement Learning. *arXiv* http://arxiv.org/abs/1812.00285
46. **Beck J.,** et al. (2025) A Tutorial on Meta-Reinforcement Learning. *Found Trends Mach Learn* **18**:224-384
47. **Derosiere G.,** Shokur S., Vassiliadis P. (2025) Reward signals in the motor cortex: from biology to neurotechnology. *Nat. Commun* **16**:1307
48. **Hosp J. A.,** LuQ A. R. (2013) Dopaminergic meso-cortical projections to M1: role in motor learning and motor cortex plasticity. *Mov Disord* **4**:145
49. **Molina-Luna K.,** et al. (2009) Dopamine in Motor Cortex Is Necessary for Skill Learning and Synaptic Plasticity. *PLOS One* **4**:e7082

50. **Ghanayim A.**, et al. (2025) VTA projections to M1 are essential for reorganization of layer 2-3 network dynamics underlying motor learning. *Nat. Commun* **16**:200

51. **Schulman J.**, Wolski F., Dhariwal P., Radford A., Klimov O. (2017) Proximal Policy Optimization Algorithms. *arXiv* https://doi.org/10.48550/arXiv.1707.06347

52. **Song S.**, et al. (2021) Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation. *J. NeuroEngineering Rehabil* **18**:126

Haith A., Yang C., Pakpoor J., Kita K. (2022) De novo motor learning of a bimanual control task over multiple days of practice. Johns Hopkins Data Research Repository. https://doi.org/10.7281/T1/QQ77OB

Peer reviews

Reviewer #1 (Public review):

Summary:

This study proposes a simple and universal reinforcement-learning framework for understanding learning in complex motor tasks. Central to the framework is a policy-gradient algorithm, in which motor commands are updated not via the gradient of the reward with respect to policy parameters, but via the gradient of the policy itself, scaled by reward information. The authors demonstrate that this scheme can reproduce learning dynamics that have been reported in previous empirical studies.

Strengths:

The key contribution of this study lies in its application of a policy-gradient algorithm to describe motor learning processes. This idea is biologically plausible, as computing the gradient of the policy with respect to its parameters is likely to be substantially easier for the nervous system than computing the gradient of the reward with respect to policy parameters. The authors present three representative examples showing that this scheme can capture several aspects of motor learning dynamics. Notably, providing such a unified description across different tasks has been difficult for conventionally proposed learning frameworks, such as supervised learning.

Weaknesses:

While this scheme is valuable in that it captures certain aspects of learning dynamics, I find that its overall significance is limited for the following reasons.

(1) The empirical results examined in this study primarily demonstrate that motor learning drives performance toward the spatial task goal while reducing variability. Given that the policies are expressed using Gaussian distributions and that their parameters (i.e., the mean and covariance matrix) are updated during learning, it is not surprising that the proposed scheme can reproduce these results by fitting the parameters to the data.

(2) The proposed framework assumes that the motor learning system relies on the gradient of the policy with respect to its parameters. However, I am not convinced that this assumption is always appropriate, because in all three empirical studies examined here, explicit spatial error information is available. In such cases, the motor learning system could, in principle, compute the gradient of the error with respect to the policy parameters directly, without relying on a policy-gradient mechanism.

(3) Most importantly, it remains unclear how the proposed scheme advances our understanding of the underlying learning mechanisms beyond providing a descriptive account of the learning process. While the framework offers a compact mathematical

description of learning dynamics, it is uncertain how it can yield novel mechanistic insights or testable predictions that distinguish it from existing learning models.

https://doi.org/10.7554/eLife.109934.1.sa1

Reviewer #2 (Public review):

Summary:

In this study, Haith applies, and to some extent extends, the theoretical framework of policy gradient (PG) and the derived REINFORCE learning rules to human motor learning. This approach is coherent because human motor skill learning is characterized by improvements in both accuracy and precision (the inverse of variance), and REINFORCE provides update rules for both the mean and the variance of the motor commands.

Weaknesses:

The mean update (equation 4) is given in task space (i.e., angle and velocity for the skittle task), but the covariance update (equation 5) is given in eigenvector space. This formulation appears to have been provided for computational convenience, as it ensures that the variances are always positive by exponentiating the eigenvalues. However, this eigenspace formulation is somewhat artificial and complex (notably the update rule for the orientation of the covariance matrix) and seems far from biological reality. A simpler alternative, suggested by the author, is to provide the full covariance matrix, including crossed terms, and derive equations to update the diagonal variance terms and the cross-terms (perhaps after a transformation to keep all elements positive if needed). This would provide a simpler and more biologically plausible update to the covariance matrix terms, in the spirit of the original REINFORCE algorithm. The author suggests that he has derived the update rule for the cross terms, so this should be relatively easy to write and update, especially for the skittle learning rules. If the author wishes to keep their rules in simulations, then the two mathematical rules could be presented in the methods or a supplementary material section.

The discussion about binary rewards and the increase in variance in previous experiments is potentially interesting. However, I do not understand why variance cannot increase with the policy-gradient RL update? Surely, equation 5 can lead to both an increase and a decrease in variance depending on the reward prediction error and the noise (for example, suppose the noise at trial i is small and leads to a smaller reward than the baseline; variance would increase). It would be interesting to see detailed simulation results for the skittle task showing changes in both mean and variance across a few consecutive trials, with both increases and decreases in reward prediction errors. These results could then be compared in simulations with those of a task with discrete binary rewards.

Generalization is a major feature of human learning, but it is not discussed or studied here. In fact, in the de novo task simulations, there can be no generalization because the values are modeled as running averages for each target rather than derived from a critic network. Can the author discuss this point and, ideally, show generalization results in simulations, say in the skittle task?

The application of the model to reproduce the Shmuelof et al. data is, at the same time, justified (because one of their main results is an improvement in precision, which Policy Gradient directly addresses) and somewhat "forced," as the author approximates curved movements with a series of straight-line movements. The author therefore needs to specify multiple via points with PG updating and a reward function that also enforces smoothness. The justification for the Guigon 2023 model seems somewhat artificial because it mainly applies to slow movements. Can the author comment and discuss alternatives that do not

require via points, drawing from the robotics literature if needed (Schaal's Dynamic Movement Primitives come to mind, for example).

Policy Gradient requires both a "noisy" and a clean "pass", making it non-biological in its simplest form. Legenstein et al. (2010) and Miconi (2017) provided biologically plausible forms for the mean update. Since Policy Gradient is proposed as a model of human motor learning, can the author discuss the biological plausibility of the proposed learning rules and possible biologically plausible extensions?

https://doi.org/10.7554/eLife.109934.1.sa0