Integrating Reward- and Error- Based Learning via Action Gradients: A Systems Theory of Cerebellar and Basal Ganglia Interactions

Michele Garibbo
University of Oxford
michele.garibbo@dpag.ox.ac.uk

Laurence Aitchison *
University of Bristol

Rui Ponte Costa *
University of Oxford

Abstract

Motor learning in the brain depends on both reward and sensory error signals, but a unified theoretical framework has yet to be established. Here, we build on reinforcement learning (RL) theory to introduce an 'action-gradient' model that integrates reward and sensory error signals into a cohesive learning process. Firstly, we demonstrate that our model can replicate a wide range of behavioral patterns observed during motor learning task, involving both rewards and sensory errors. Secondly, we propose a systems-level implementation of this model in the brain, involving close interactions between the cerebellum and basal ganglia. We show the model accounts for classical observations in both the cerebellar and basal ganglia neuroscience literature, while providing novel experimental predictions. In summary, our work presents a systems level framework for understanding how reward and sensory error signals may drive motor learning in the brain.

1 Introduction

Motor commands typically result in two types of feedback: a sensory outcome reflected in the activity of sensory areas (e.g., vision and proprioception), and a reward, manifested as a subjective assessment of the utility of these sensations. In the neuroscience literature, sensory outcomes and rewards have been studied in relation to two distinct (learning) frameworks: reward-based (RBL) and error-based (EBL) learning [70, 41, 101]. RBL aims to explain learning from rewards and has been related to stochastic policy gradient computations within the basal ganglia, in line with actor-critic algorithms from the reinforcement learning (RL) literature [67, 43, 53, 5, 4]. Error-based learning, EBL, aims to describe learning from sensory outcomes and has been associated with the cerebellum and supervised learning [22, 50, 25]. Traditionally, the cerebellum is believed to provide an internal model of the world (e.g., forward or inverse model predictions), which learns based on (supervisory) error feedback from the inferior olives [100, 56, 40, 3].

Crucially, RBL and EBL have typically been regarded as separate learning systems [e.g., 101, 22, 41, 21], in line with the traditional belief the cerebellum and the basal ganglia form independent channels in the thalamus [59, 35, 61]. This belief has led to the implicit assumption rewards (RBL) and sensory errors/outcomes (EBL) drive learning of separate policies (motor commands) in the brain (e.g., one policy stored in the basal ganglia and another adaptive policy stored in cerebellum) [e.g., 41, 90, 45, 4]. However, there is no clear explanation to how the brain may coordinate learning

[*] These authors contributed equally to this work

Preprint. Under review.

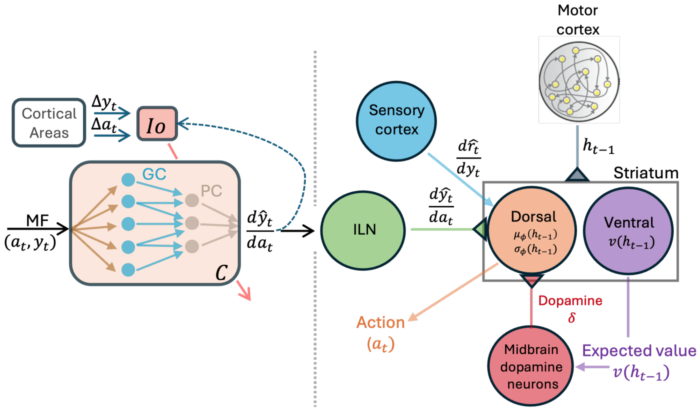

Figure 1: [System-level circuit] ***left-only****) Representation of the cerebellar network as a 1 hidden-layer feedforward neural network [e.g., 8, 71]. The mossy fibers (MF) convey the cerebellar input, encoding the motor commands, a and the corresponding (sensory) outcomes, y, at the end of each trial/step, t. Granule cells (GC) represent the hidden layer computations. Finally, Purkinje cells (PC) represent the output layer, providing the action-to-outcome gradient predictions, $\frac{d\hat{y}}{da}$. This cerebellar network learns based on the inferior olive (Io) feedback, conveying a mismatch between the (current) cerebellar output and observed changes in (sensory) outcomes (i.e., Δy) relative to the changes in motor commands (i.e., Δa) [i.e., exploiting a finite difference gradient estimate 86] (see section A.6 for more details).* ***entire****) System level implementation of the proposed action-gradient model within a cerebellar basal ganglia circuit [6] (see section 4). The motor cortex provides latent (motor) representations, h_{t-1}, to the striatum, while sensory cortical areas provide estimates of the 'directed' sensory error, $\frac{d\hat{r}}{dy}$. Finally, dopaminergic projections from midbrain dopamine neurons convey RPEs, δ_t to the striatum. RPes are computed based on expected rewards, $v(h_{t-1})$ conveyed to midbrain dopamine neurons by the ventral striatum. Io: inferior olive, C: cerebellum ILN: intralaminar nuclei of the thalamus.*

across separate RBL and EBL polices to solve the same underlying task. For instance, it is not trivial how information acquired based on rewards generalizes to the error-driven policy and vice-versa or, how the brain is able to maintain the two policy synchronized (see Appendix A.2). It is also not clear why the brain would benefit from storing and updating two separate policies to perform the same underlying task. This is despite clear evidence the brain is able to smoothly learn from combinations of sensory and reward signals within the same task [e.g., 41, 26, 70, 66]. Moreover, it is now evident the basal ganglia and the cerebellum have much tighter anatomical and functional connections, discrediting the (traditional) belief that the cerebellum and the basal ganglia form independent learning channels [e.g., 39, 35, 102, 7, 12, 13, 61, 6, 90]. For instance, several studies have found strong disynaptic projections from the cerebellum to the dorsal striatum (i.e., putamen and caudate) via the intralaminar nuclei of the thalamus (ILN) [e.g., 39, 35, 102, 13]. Additionally, the cerebellum seems to influence (dorsal) striatal activity as well as plasticity with observable behavioural consequences [13, 73, 102, 6], highlighting a potential role of the cerebellum in driving learning in the dorsal striatum. Therefore, we believe the brain may more plausibly integrate reward (RBL) and sensory error (EBL) signals into a unified teaching signal, rather than each driving learning of separate policies and, this process may reflect tight basal ganglia and cerebellar interactions.

To support this proposal, we introduce a simple computational model that allows to integrate RPEs (RBL) and sensory errors (EBL) into unified teaching signals to drive learning. To do so, we build on previous theoretical work highlighting that RBL and EBL respectively transform RPEs and sensory errors into 'action gradients' (i.e., gradients of a scalar reward/error relative to the actions) [27]. This shared 'action gradient' representation allows the brain to combine RBL and EBL feedback in the same (action gradient) space, which would not be possible by directly considering RPEs and sensory errors. The resulting weighted combination of RBL and EBL feedback in action gradient space provides a unified teaching signal, which the brain may exploit to drive synaptic plasticity of motor commands (i.e., policy) in response to reward and sensory (error) signals. To support this, we show the proposed action gradient model is able to reproduce a broad range of behavioural observations on

RBL and EBL interactions [e.g., 41, 19, 20, 99]. Next, we provide a system-level implementation relating the (action gradient) model to the mounting evidence of tight basal ganglia and cerebellar interactions during learning [e.g., 13, 73, 102, 7, 12, 61] (see Fig. 1). We introduce a cerebellar network, whose predictions allow to compute the EBL action gradient, while we assume dopamine projections conveying reward prediction errors (RPEs) allow to compute the RBL action gradient [i.e., in line with classical RBL models 67, 43, 53, 4]. We show this model combining cerebellar-dependent (EBL) and dopamine-dependent (RBL) action gradients account for classical observations in both the cerebellar and basal ganglia neuroscience literature, while providing novel predictions on how the two structures may interact.

2 Background

Here we briefly describe how RBL and EBL can be characterised by different gradient updates (i.e., synaptic learning rules). To better frame these learning rules, we borrow the terminology of the RL literature. We represent motor commands as actions, a, sampled from a policy, π_ϕ, which is encoded by synaptic weights ϕ. The (policy) synaptic weights, ϕ, must be updated to to maximize some (scalar) reward/error signal, r_t (i.e., representing successful motor learning), across each time step or trial, t. In line with previous work [53, 27], we assume a Gaussian policy generates the actions, $\pi_\phi(a_t \mid s_t) = \mathcal{N}(a_t \mid \mu_\phi(s_t), \Sigma_\phi(s_t) + \Sigma_{\text{fixed}})$, where the mean, μ_ϕ, and (diagonal) covariance, Σ_ϕ, are the outputs of some motor area with synaptic weights, ϕ (i.e., the synaptic weights of the area encoding the policy), which receives some cue/state, s_t, as input at each trial/step (see Appendix A.3 for further details on the employed policy parametrizations). The term Σ_{fixed} is used to introduce some constant irreducible motor noise [i.e., a small amount of motor noise is always present in human behaviours 20]. Below, we present how RBL and EBL adapt the (synaptic) weights of the (Gaussian) policy mean, μ_ϕ to maximise the reward/error signal, r (see section A.7 for the Gaussian variance update, σ^2_ϕ). We do so in terms of ballistic (feedforward) motor movements, since this is most relevant for the motor adaptation paradigms we model subsequently. Therefore, each action, a_t, results in a single reward, r_t and a terminal (sensory) outcome, y_t (i.e., equivalent to terminal states in RL jargon) for each trial t. Nevertheless, these computations can easily be extended to sequential (feedback) tasks (e.g., by introducing a value function). For simplicity, we present all gradients computations in terms of 1-dimensional actions performed in a 1-dimensional sensory space, although this can easily be extended to higher dimensional cases (e.g., line drawing task).

Reward-based Learning. Traditionally, RBL is associated to stochastic policy gradient updates [67, 5, 53, 43],

$$\mathrm{E}_{\pi_\phi}[\Delta\phi_t]^{\mathrm{RBL}} \propto \mathrm{E}_{\pi_\phi}\left[\nabla_\phi \log(\pi_\phi(a_t \mid s_t)\delta_t\right] \tag{1}$$

where $\Delta\phi_t$ denotes the estimated change to the synaptic weights responsible for the (mean) actions (i.e., the policy mean), while $\delta_t = r_t - v_\theta(s_t)$ encodes the RPEs, for some observed reward signal, r_t, and the corresponding expected reward, $v_\theta(s_t)$, given some sensory cue/state, s_t. Note, here t could denote the time step or the trial number. RBL has closely been related to the function of the Basal Ganglia. The actor, π_ϕ, is often associated to the dorsal striatum, which appears key for action selection [43, 69]. Conversely, the ventral striatum is thought to play the role of a critic, v_θ, (i.e., providing expected rewards) given its more prominent role in value-estimation rather than action selection [63, 67]. Finally, striatal dopamine projections are thought to convey RPEs, δ_t, needed to learn both the actor and the critic [43]. This is in line with the extensive evidence of striatal dopamine activity increases in response to unexpected rewards/cues and suppresses in face of unexpected lack of rewards [83, 63, 37].

Error-based learning. EBL is typically described by the following synaptic weight update [30, 1, 44, 27],

$$\mathrm{E}_{\pi_\phi}[\Delta\phi_t]^{\mathrm{EBL}} \propto \mathrm{E}_{\pi_\phi}\left[\frac{dr_t}{dy_t}\frac{dy_t}{da_t}\frac{d\mu_\phi}{d\phi_t}\right] \tag{2}$$

where $\Delta\phi_t$ again denotes the estimated change to the synaptic weights at the trial t, the term y_t represents the terminal (sensory) outcome in response to the action, a_t, while r_t represents some (scalar) error/reward signal (often denoted by e). Again, in RL jargon, the (sensory) outcome, y_t, can be thought as the terminal state, s_T and each trial, t can be thought as an entire episode. In motor learning paradigms, this error/reward signal is typically computed as the squared distance between the terminal, y_t, and the desired, y^*, (sensory) outcomes at any given trial, t (i.e., $r_t = (y(a_t) - y^*)^2$) (e.g., the distance from the goal state after taking action, a_t). In eq. 2, the operator $\frac{dr}{dy}$ represents a 'directed' sensory error, encoding how to change the sensory outcome, y, to reduces the scalar error, r (e.g., signaling to throw more to the right to hit a target in a dart game). The operator $\frac{dy}{da}$ provides an action-to-outcome 'interface', enabling to map the (directed) sensory errors, $\frac{dr}{dy}$, to motor changes (e.g., instructing how to change the action to throw more to the right). Finally, $\frac{d\mu_\phi}{d\phi}$ determines how to change the synaptic weights to change the (mean) action in the desired direction. Note, to correctly estimate the product $\frac{dr}{da}\frac{d\mu_\phi}{d\phi}$, the reparameterization trick can be used in case the relation between actions, a, and the policy parameter, μ_ϕ, is stochastic [46] (e.g., a with Gaussian policy).

3 Action-gradient model: Unifying reward- and error- based learning

For a Gaussian policy, the RBL update (eq. 1) can be factorised into two terms, 1) an 'action gradient' term, $\frac{dr}{da}$, which should be regarded as a teaching signal, instructing how the current action should change to improve performance and, 2) a gradient term, $\frac{d\mu_\phi}{d\phi}$, which drives plasticity of the synaptic weights, ϕ, responsible for the (mean) action [see 27, 53],

$$\mathrm{E}_{\pi_\phi}[\Delta\phi_t]^{\mathrm{RBL}} \propto \frac{dr_t}{da_t}\frac{d\mu_\phi}{d\phi_t} = \delta(a_t - \mu_\phi(s_t))\frac{d\mu_\phi}{d\phi_t} \tag{3}$$

where again δ denotes the RPE for the action, a. Crucially, the EBL update (i.e., eq. 2) takes a similar form, where the teaching signal (i.e., the action gradient term) is composed of two sub-terms, $\frac{d\hat{r}}{da} = \frac{d\hat{r}}{dy}\frac{d\hat{y}}{da}$, to model the explicit dependency of the update on the sensory outcomes y. This similarity allows us to perform a weighted combination of the RBL and EBL teaching signals in action gradient space to drive synaptic plasticity of a shared motor area, ϕ (i.e., encoding the policy),

$$\mathrm{E}_{\pi_\phi}[\Delta\phi_t]^{\mathrm{Mixed}} \propto \mathrm{E}_{\pi_\phi}\underbrace{\left[\overbrace{\frac{dr_t}{dy_t}\frac{dy_t}{da_t}}^{\substack{\text{EBL}\\ \text{action gradient}}}\beta + \overbrace{\delta(a_t - \mu_\phi(s_t))}^{\text{RBL action gradient}}(1-\beta)\right]}_{\text{Mixed action gradient}}\overbrace{\frac{d\mu_\phi}{d\phi_t}}^{\text{Synaptic plasticity}} \tag{4}$$

where $\beta \in [0, 1]$ controls the mixture between the two action gradient terms (i.e., teaching signals), encoding the uncertainty over the EBL action gradient term (i.e., the higher the uncertainty the closer should β be to zero). Crucially, the uncertainty over the EBL action gradient should directly relate to the uncertainty over the sensory outcome, y (i.e., the quality of the sensory feedback). This is because the EBL action gradient directly depends on the sensory outcome, y. At the extreme case where no sensory information is provided, but only a binary success/failure reward signal, the EBL action gradient cannot be estimated, implying learning should exclusively be driven by the RBL action gradient (i.e., $\beta = 0$) [e.g., 88, 95]. At the other extreme, where 'perfect' sensory information is available, the EBL action gradient should provide the optimal learning strategy (i.e., $\beta = 1$) [e.g., see 27, 32, 15]. Beyond these two extremes, the brain seems to rely on a mixture of RBL and EBL, where the amount of EBL contribution negatively relates to the uncertainty over the sensory feedback [e.g., 99, 10, 91, 41]. This represents a key prediction of the proposed (action gradient) model, which uses the β hyper-parameter to model a RBL-EBL mixture in terms of the corresponding action gradients. In summary, the update in eq. (4) proposes that the brain unifies EBL and RBL at the level of teaching signals (i.e., action gradients) to drive synaptic plasticity of shared motor areas, ϕ (i.e., a shared policy) instead of each driving learning at separate motor areas (i.e., at separate RBL and EBL policies). It is worth stressing this update (eq. 4) is consistent with a 3 factor learning rule of synaptic plasticity [51], where the "mixed-action gradient" can be thought as an error signal. In Appendix A.7, we provide further information on how each component of this Mixed action gradient is computed in practice.

4 System-level circuit

Here, we explore how our action-gradient mechanism might provide a framework to explain the mounting evidence of tight cerebellar and basal-ganglia interactions [13, 73, 102, 7, 12, 61]. In line with the classical view of basal ganglia [43, 4, 67], we assume that the dorsal and ventral striatum, respectively, play the role of policy and critic, while midbrain dopaminergic projections convey RPEs, δ, to both dorsal and ventral striatum [83, 67]. These dopamine-dependent RPEs appear to drive cortico-striatal plasticity at the dorsal striatum, in line with the RBL action gradient [67, 77, 43, 11] (note, the action term $(a - \mu_\phi)$ is typically assumed to be available at the dorsal striatum post-synapses 53). Cortico-striatal synapses may convey (latent) representations, h, of actions (e.g., "motor plans") from the primary motor cortex (M1) to the striatum [22, 21, 58, 29]. In turn, the dorsal striatum learns to select the optimal actions (i.e., acting as a policy) based on the dopamine-dependent RPEs (while the ventral striatum learns the expected value of each plan) [4, 43].

Crucially, our proposed action gradient model may extend this view, by integrating well-documented cerebellar-dependent signals into this learning process [e.g., 6, 13, 73, 102]. The model (eq. 4) predicts dorsal striatal plasticity (i.e., the policy synaptic weights, ϕ) should not only be driven by dopamine, but also by cerebellar-dependent signals, reflecting a combination of dopamine-dependent RBL and cerebellar-dependent EBL feedback (i.e., action gradients). Specifically, the EBL action gradient is composed of two quantities, an action-to-outcome gradient, $\frac{dy}{da}$, and a 'directed' sensory error $\frac{dr}{dy}$ (see eq. 4). We propose the cerebellum learns to predict the action-to-outcome gradient, $\frac{dy}{da}$, in line with the assumption this quantity must be learned from experience [see 1, 44, 38, 27]. We denote the cerebellar predicted action-to-outcome gradient as $\frac{d\hat{y}}{da}$, where $\frac{d\hat{y}}{da} \approx \frac{dy}{da}$. We think the cerebellum is well placed to learn the action-to-outcome gradient, $\frac{dy}{da}$, given its key role in encoding internal models of the world [21, 100, 50, 71, 8] and its importance to EBL [50, 94, 92, 42]. Conversely, we assume (sensory) cortical areas estimate the 'directed' sensory error $\frac{dr}{dy}$, depending on the sensory modality in which the task sensory outcomes are perceived (e.g., in case of abstract errors/rewards, this quantity may be computed by the prefrontal cortex). We denote the (cortical) prediction of 'directed' sensory error as $\frac{d\hat{r}}{dy}$, where $\frac{d\hat{r}}{dy} \approx \frac{dr}{dy}$. Both the cerebellar and the cortical predictions converge to the dorsal striatum, the former via targeted disynaptic projections going through the intralaminar nuclei of the thalamus (ILN) [e.g., 102, 73, 6], while the latter via the well-known cortico-striatal projections [72]. The two predictions provide the necessary information to compute the EBL action gradient and influence cortico-striatal plasticity together with the dopamine-dependent RBL action gradient (i.e., in line with eq. 4). In support to this proposal, cerebellar projections seem to target (dorsal striatal) medium spiny neurons, which also possess D1 and D2 dopamine receptors [102] (i.e., enabling cerebellar and dopamine feedback to converge within the same neurons). In summary, this circuit (see Fig. 1) could drive motor learning in the dorsal striatum (policy) based on a mixture of cerebellar- and dopamine- dependent teaching signals, reflecting RBL and EBL mechanisms, as described by the action-gradient framework in eq. (4). Note, our action-gradient framework assumes some mechanism to regulate the dopamine-dependent (RBL) action gradient in relation to the cerebellar-dependent (EBL) action gradient at the dorsal striatum (i.e., the β-weighting mechanism in eq. 4). We speculate this mechanism could take place implicitly, through (synaptic) competition mechanisms driven by the magnitude of each gradient or, explicitly, with the cerebellum directly controlling midbrain dopamine release at the striatum, as some evidence indicates [64, 65, 97, 103].

5 Results

5.1 Action-gradient model predictions of behavioural data

In this section, we show our action-gradient model (section 3) is able to replicate a series of behavioural findings on RBL and EBL interactions during targeted arm reaching experiments. These experiments are primarily based on the seminal paper by Izawa and Shadmehr [41] on RBL and EBL interactions in humans. We model the RBL and EBL conditions by respectively setting $\beta = 0$ and $\beta = 1$ in our action-gradient model [respectively denoted as RWD and ERR in the original paper 41]. This is because $\beta = 0$ implies synaptic plasticity is exclusively driven by RPEs (i.e., RBL), while $\beta = 1$ implies synaptic plasticity is exclusively driven by sensory errors (i.e., EBL) (see eq. 4). Conversely, we found $\beta = 0.4$ could best predict Izawa and Shadmehr [41]'s result on the 'Mixed' condition, where both rewards and sensory errors are likely driving learning (i.e., denoted as EPE in the original paper). The task set-up is the same targeted reaching task as described by Izawa and Shadmehr [41], where the agent needs to adapt to a visuomotor perturbation, which is gradually increased over a series of trials (see Appendix A.4.1 for further details).

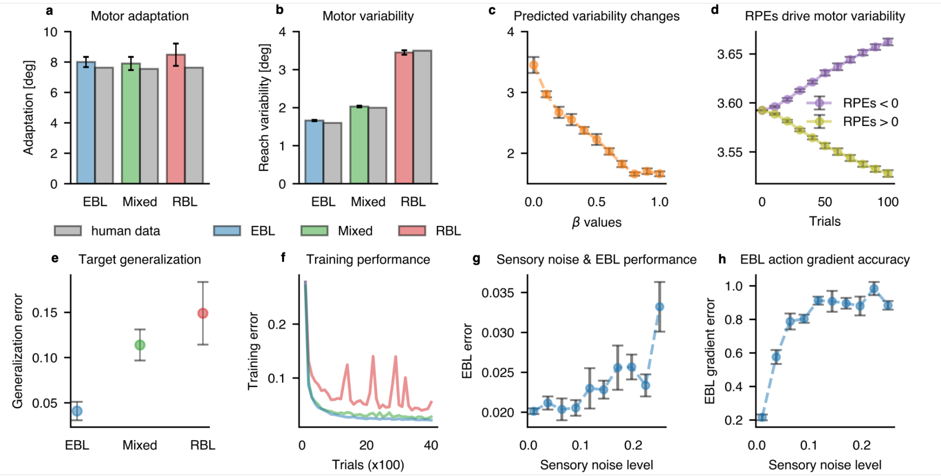

Figure 2: ***a-b)*** *performance comparison between our action-gradient model and human experimental data during EBL (blue), RBL (red) and a mixture of the two (green) (the human data were re-plotted from [41] using WebPlotDigitizer [79]).* ***c)*** *model predicted change in motor variability as we move from purely RBL teaching signals ($\beta = 0$) to higher contributions from the EBL teaching signals (i.e., higher β-values)****d)*** *model predicted change in motor variability during RBL in response to either negative (purple) or positive (yellow) RPEs.* ***e-f)*** *action-gradient model's generalisation performance (**e**) as well as training error (**f**) across EBL, RBL and a mixture of the two.* ***g-h)*** *EBL performance (**g**) as well as EBL action gradient accuracy (**h**) across different levels of sensory noise. Error bars are computed based on performances on five random seeds.*

Motor variability changes Fig. 2b shows our action-gradient model is able to predict human motor variability changes across RBL, EBL and a mixture of the two, while correctly predicting the same amount of (human) motor adaptation across the 3 conditions (Fig. 2a). Note, while the model predicts full adaptation to the (8-degree) perturbations across the 3 conditions, humans seem to slightly under-adapt across all the conditions. This is may be due to humans having a slight bias towards a side of the target [e.g., 30]. Crucially, from the RL literature, we know the RBL action gradient to be higher variance (i.e., noisier) than the equivalent EBL action gradient (known as model-based policy gradient in RL jargon) [32, 27]. Therefore, our action-gradient model suggests changes in human motor variability across RBL-EBL conditions may be explained by the variability of the different action gradient estimates (i.e., the variability of RBL and EBL teaching signals as well as any mixture of the two). The model also allows to obtain continuous predictions on how quickly motor variability may decrease as we move from purely RBL teaching signals (i.e., $\beta = 0$) to higher contributions from the EBL teaching signals (i.e., higher β-values, see Fig. 2c).

RPEs drive motor variability During RBL, human motor variability has been shown to increase in response to negative RPEs, while decreasing in response to positive RPEs, [i.e., efficiently controlling the amount of exploratory behaviours through motor variability 19, 20]. Crucially, our action-gradient model is able to replicate these changes in (human) motor variability in response to different RPEs. To demonstrate this, we show the model predicts the action variability to increase in response to negative RPEs, while decreasing in response to positive RPEs (see Fig 2c) (i.e., we fixed the RPEs to always be negative or always be positive, and set $\beta = 0$ to model RBL only). This patterns follows directly from how the RBL action gradient changes the action variance in response to positive and negative RPEs (see Appendix A.7).

Target Generalisation Fig. 2e shows our action-gradient model predicts RBL to achieve poorer generalisation to novel targets compared to EBL. The mixture of RBL and EBL teaching signals (i.e., action gradients) achieves intermediate generalisation performance (i.e., the generalisation error is in between the RBL and EBL conditions). Crucially, the same generalisation pattern is observed in humans across the three conditions [41]. To obtain the mixture of RBL and EBL teaching signals, we used the same β-value as for the human motor variability results described above (i.e., $\beta = 0.4$). This accounts for the fact that both the human motor variability and generalisation results were based on the same task set-up with the same quality of sensory information [i.e., 41]. Fig. 2f shows that the poorer generalisation performance of RBL arises from the fact that during training RBL converges to a poorer policy compared to EBL and the Mixed condition (i.e, in terms of training errors). Finally, we wanted to get some insights on why human subjects tend to rely less on EBL adaptation in the presence of poorer sensory feedback [99, 10, 91, 41]. Fig. 2h show the EBL action gradient becomes increasingly less accurate as the sensory noise increases, impairing performance (Fig. 2g) [e.g., 99]. This was measured as the l2-norm between the ground truth gradient and the estimated EBL gradients across different levels of sensory noise. We explore this further in the "Optimal cerebellar and dopamine learning combinations" section below.

5.2 Action-gradient model of cerebellar and basal ganglia interactions

Here, we want to explore the (action-gradient) model ability to explain well-known cerebellar and basal ganglia observation during motor learning, while exploiting the model to make novel experimental predictions. Following the system-level circuit described in section 4, we assume dopaminergic (DA) and cerebellar (CB) signals respectively enable to compute the RBL and EBL action gradients. We use these two signals to drive policy learning (e.g., at the dorsal striatum) in line with the action gradient computations proposed in eq. 4. We model the cerebellum with a 1 hidden layer feedforward neural networks, which learns to predict the action-to-outcome gradient estimates, $\frac{dy}{da}$ form the current action, a and the (observed) sensory outcome, y [e.g., 21] (see Appendix A.6 for more details). These cerebellar predictions are then used to estimate the EBL action gradient together with (cortical) estimates of the 'directed' sensory errors, denoted as $\frac{d\hat{r}}{dy}$, which we assume to be readily available from sensory information. The choice of architecture for the cerebellar module is consistent with previous work [e.g., 8, 71] and, provides a simplified model for the highly uniform structure of the cerebellum [85, 93]. Conversely, we assume dopamine conveys RPEs, δ, allowing to compute the RBL action gradient [43, 53, 4, 67]. Based classical cerebellar and basal ganglia studies of motor learning [80, 92, 42, 33, 28, 82], we test our model on several motor tasks, involving a (non-linear) kinematic model of the arm [17] (see Appendix A.5).

Visuomotor rotation task Here, we test our (action-gradient) model in a classical cerebellar-dependent adaptation task, in which an agent needs to adapt to a visuomotor rotation that is suddenly applied during targeted arm-reaches [e.g., 50, 57, 42, 75]. We employ the same task set-up as Tseng et al. [92], for healthy controls vs cerebellar patients (see Appendix A.4.2 for further details). Fig. 3a shows the model accurately reproduces the difference in adaptation between healthy controls and cerebellar patients, doing so exclusively in terms of how the RBL and EBL action gradient estimates drive performance. Specifically, the model predicts healthy controls' performances are best captured when the cerebellar-dependent EBL action gradient drive learning fully (i.e., CB-driven, $\beta = 1$ in eq. 4), achieving lower residual errors (i.e., optimal strategy). Conversely, the model predicts cerebellar patients' poorer performances are best captured when the dopamine-dependent RBL action gradient drives learning fully (i.e., DA-driven, $\beta = 0$ in eq. 4). This finding suggests cerebellar patients' deficits may prevent them from exploiting the (optimal) EBL action gradient, forcing them to rely on the (sub-optimal) RBL action gradient (i.e., DA-driven), resulting in higher residual errors. On the flip side, this finding implies cerebellar patients may still be able to learn from the dopamine-dependent RBL action gradient, as some evidence seems to suggest [e.g., 88, 89]. Crucially, our model also allows us make predictions on how adaptation performance may change as the contribution of the cerebellar-dependent (EBL) action gradient increases over the contribution of the dopamine-dependent (RBL) action gradient (i.e., by increasing β in our model). Fig. 3b shows the residual error should decrease smoothly as the cerebellar contribution increases, peaking when the cerebellum-dependent (EBL) action gradient drives performance fully. This model prediction may explain why cerebellar symptom severity appears to positively correlate with adaptation errors [92], suggesting less-severe cerebellar patients may still be able to exploit some of the contribution of the cerebellar-dependent (EBL) action gradient instead of having to rely on dopamine feedback exclusively.

Visuomotor reversal task Here, we show our action gradient model is able to reproduce human control and basal ganglia patients (i.e., Parkinson's and Huntington's) data on a visuomotor reversal task based on Gutierrez-Garralda et al. [28] (see Appendix A.4.2 for further details). Unlike the previous (rotation) task, visuomotor reversal tasks reverse sensory (visual) feedback in a way that sensory errors become detrimental to learning [e.g., worsening performance with learning, 30, 52, 96]. For this reason, successful adaptation to visuomotor reversal tasks should not be driven by sensory errors (EBL), but by reward signals (RBL) [e.g., 28, 82, 27] (i.e., providing an opposite test to the

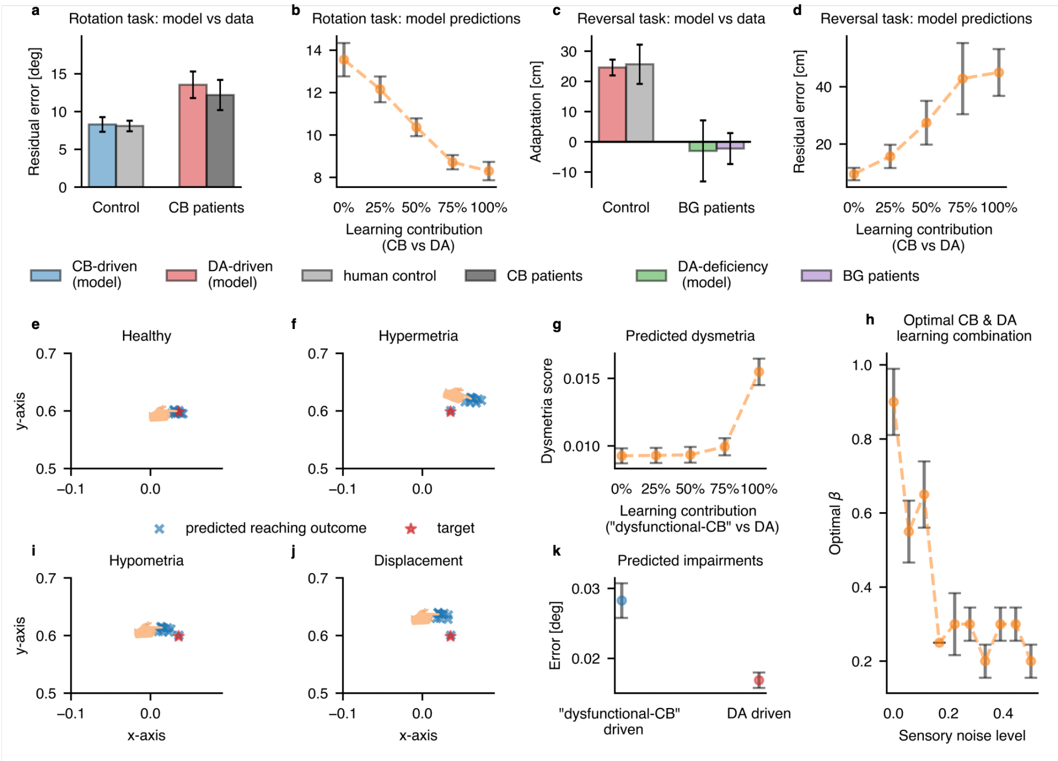

Figure 3: ***a-b**) Sensorimotor rotation task. **c-d**) Sensorimotor reversal task. **e-g,i-k**) Model predictions on a standard reaching task with a perturbed cerebellar component. **h**) Optimal combination of RBL and EBL action gradients across sensory noise levels. In (**a**) the human control and CB patients were re-plotted from [92], while in (**b**) the human control and BG patients were re-plotted from [28], using WebPlotDigitizer [79]. CB: cerebellum, BG: basal ganglia. Error bars are computed based on performances on five random seeds.*

visuomotor rotation task, where cerebellum-dependent EBL appeared to be the optimal strategy). Fig. 3c shows our action gradient model accurately reproduces human controls' successful adaptation to the reversal task [i.e., 28], by exploiting the dopamine-dependent RBL action gradient (i.e., DA-driven, $\beta = 0$). Crucially, we found the model was able to reproduce the poor performance of basal ganglia patients in the reversal task, using $\beta = 0.7$ (see Fig. 3c). This high β value (i.e., the max is 1) suggests learning in basal ganglia patients was only marginally driven by the (optimal) dopamine-dependent RBL action gradient (i.e., DA-deficiency), unlike healthy controls who fully exploited the (optimal) dopamine-dependent RBL action gradient to solve the task. This finding is strikingly consistent with the well-know dopamine deficiencies in Parkinson's patients [16, 76, 81] and in later-stage Huntington's patients [14]. Finally, the model predicts residual errors should positively correlate with dopamine deficiency (i.e., symptoms severity) in visuomotor reversal tasks (see Fig. 3d). To our knowledge, this prediction remains to be tested.

Dysfunctional cerebellar predictions during reaching In the visuomotor rotation task, we found cerebellar patients' deficits were best describe by a failure to access the (optimal) cerebellar-dependent EBL action gradient (i.e., being limited to the sub-optimal dopamine-dependent RBL action gradient). Nevertheless, our action gradient model also allows us to make predictions on what happens when the EBL action gradient is still accessible for learning, but it is computed based on erroneous cerebellar predictions (i.e., erroneous action-to-outcome gradient estimates, $\frac{d\hat{y}}{da} \neq \frac{dy}{da}$). For instance, this could happen when the cerebellum is still present, but it is sending erroneous predictions due to anatomical or functional impairments. In Fig. 3, we denote this condition as "dysfunctional-CB" driven and investigate it in a standard reaching task, where the agent had to reach for a target under normal conditions. We found altering specific components of the cerebellar module predictions, $\frac{d\hat{y}}{da}$, led to very specific motor deficits (see Fig. 3e,f,i,j) (see Appendix A.4.3 for further details). Strikingly, all the deficits seem consistent with well-known dysmetria symptoms in cerebellar patients [e.g., hypermetria, hypometria 33, 55, 34]. Finally, the model predicts dysmetria symptoms should improve (i.e., achieving lower dysmetria scores and reaching errors) if learning relies on dopamine at the expenses of the the dysfunctional cerebellar contribution (dysfunctional-CB). Therefore, cerebellar patients' dysmetria symptoms may improve by purposely engaging in reward-driven learning regimes (see Fig. 3g,k). Although this key prediction requires testing, some evidence suggests this may be the case [e.g., 89], potentially opening the way for novel rehabilitation interventions.

Optimal cerebellar and dopamine learning combinations A key prediction of our action gradient model is that the brain should rely on a (β-) weighted combination of the cerebellar-dependent (EBL) and the dopamine-dependent (RBL) action gradients, depending on the amount of uncertainly over the sensory outcome, y (i.e., the higher the uncertainty, the lower reliance on cerebellar-dependent EBL). To probe this prediction, we investigate which RBL-EBL gradient combinations lead to optimal learning performance across different levels of noise over the sensory outcome, paired with (noise-free) reward signals (i.e., in terms of the optimal β combination of the RBL and EBL action gradients). The exact experimental details are reported in Appendix A.4.4. Fig. 3h shows the optimal action gradient should reflect a combination of both the cerebellar-dependent (EBL) and the dopamine-dependent (RBL) action gradients across most levels of sensory noises (i.e., $0 <$ Optimal $\beta < 1$ for most sensory noise levels). Crucially, as the sensory noise increases, the contribution of the EBL action gradient should rapidly decrease in favour of higher contributions from the RBL action gradient. This is because as the sensory (outcome) noise increases, the accuracy of the EBL action gradient decreases, but this inaccuracy can be mitigated by the contribution of the RBL action gradient. In the 'real' world sensory information is encountered with different degrees of noise, implying the brain may greatly benefit from combining the two actions gradients, reflecting close basal ganglia and cerebellar interactions [e.g., 13, 73, 6, 102]. This prediction should be tested in a motor learning task where the quality of the sensory outcome is varied continuously together with (noise-free) reward signals. As the the quality of the sensory outcome decreases learning should move from cerebellar-dependent towards dopamine-dependent, reflecting an increase in motor variability.

6 Discussion

The paper introduces a novel computational framework to explain how the brain may combine sensory (EBL) and reward (RBL) signals into a unified learning process. We relate this model to a system-level circuit, which may help understand the mounting evidence of tight the cerebellar and basal ganglia interactions [13, 73, 102, 7, 12, 61]. We show the model reproduces classical observations in both the cerebellar and basal ganglia neuroscience literature [e.g., 92, 28, 33, 34], while providing novel predictions on how the two structures may interact. In particular, the model predicts the dorsal striatum (policy) may not only learn based on the well-known dopaminergic projections (RBL) [4, 43, 53], but also from cerebellar feedback combined with (cortical) sensory information (EBL) [e.g., 13]. Although this prediction requires precise testing, Chen et al. [13] showed cerebellar stimulation affect dorsal striatal plasticity with a very short latency, which may be consistent with a trial-to-trial learning mechanism. Additionally, we show perturbing the cerebellar predictions within our model leads to deficits that are strikingly consistent with classical cerebellar motor deficits [e.g., hypermetria, hypometria 33, 55, 34].

Our action gradient model may explain the roles of cerebellar and dopamine feedback beyond just policy learning in the dorsal striatum. Depending on the task, this model could apply to other areas, encoding a policy and receiving dopaminergic as well as cerebellar projections together with (cortical) sensory feedback. For instance, for higher-level motor function, the policy may be encoded in the motor cortex [e.g. 48, 68, 54, 2] and for some cognitive tasks, the policy may be encoded in the prefrontal cortex [9]. Therefore, cerebellar and dopamine feedback may respectively convey the EBL and RBL action gradients to drive policy learning at those areas in line with our proposed action gradient framework (i.e., eq. 4). In this regard, cerebellar and dopamine feedback have been related to synaptic plasticity at different cortical areas, including the primary motor cortex [dopamine-driven plasticity, 36, 62, 78][cerebellar-driven plasticity, 74, 31, 47] and the prefrontal cortex [dopamine-driven plasticity, 84, 24, 18] [cerebellar-driven plasticity, 60, 98, 23].

A main limitation of the present paper is that we replicate only behavioural, rather than also neural data. This limitation primarily arises due to the scarcity of neural data on basal ganglia and cerebellar interactions during learning. As we briefly touch below, the model makes clear predictions on these interactions, which should drive future experimental recordings to test the proposed model. A second limitation is treating β, which describes the relative importance of the RBL and EBL action gradients, as a hyperparameter. Although the primary scope of the paper is to show that a weighting combination of the two action gradients can explain how the brain lean from sensory and reward signals, future work should unpack the exact mechanism enabling the brain to do so.

References

[1] MN Abdelghani, Timothy P Lillicrap, and Douglas B Tweed. Sensitivity derivatives for flexible sensorimotor learning. *NeuralComp*, 20(8):2085–2111, 2008.

[2] Robert Ajemian, Andrea Green, Daniel Bullock, Lauren Sergio, John Kalaska, and Stephen Grossberg. Assessing the function of motor cortex: single-neuron models of how neural response is modulated by limb biomechanics. *Neuron*, 58(3):414–428, 2008.

[3] James S Albus. A theory of cerebellar function. *Mathematical biosciences*, 10(1-2):25–61, 1971.

[4] Andrew G Barto. Adaptive critics and the basal ganglia. 1995.

[5] Daniel Bennett, Yael Niv, and Angela J Langdon. Value-free reinforcement learning: policy optimization as a minimal model of operant behavior. *Current Opinion in Behavioral Sciences*, 41:114–121, 2021.

[6] Andreea C Bostan and Peter L Strick. The basal ganglia and the cerebellum: nodes in an integrated network. *NatRevNeuro*, 19(6):338–350, 2018.

[7] Andreea C Bostan, Richard P Dum, and Peter L Strick. The basal ganglia communicate with the cerebellum. *Proceedings of the national academy of sciences*, 107(18):8452–8456, 2010.

[8] Ellen Boven, Joseph Pemberton, Paul Chadderton, Richard Apps, and Rui Ponte Costa. Cerebro-cerebellar networks facilitate learning through feedback decoupling. *Nature Communications*, 14(1):51, 2023.

[9] Y Broche-Perez, LF Herrera Jimenez, and E Omar-Martinez. Neural substrates of decision-making. *Neurologia (English Edition)*, 31(5):319–325, 2016.

[10] Johannes Burge, Marc O Ernst, and Martin S Banks. The statistical determinants of adaptation rate in human reaching. *Journal of vision*, 8(4):20–20, 2008.

[11] Paolo Calabresi, Barbara Picconi, Alessandro Tozzi, and Massimiliano Di Filippo. Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in neurosciences*, 30(5): 211–219, 2007.

[12] Daniele Caligiore, Giovanni Pezzulo, Gianluca Baldassarre, Andreea C Bostan, Peter L Strick, Kenji Doya, Rick C Helmich, Michiel Dirkx, James Houk, Henrik Jörntell, et al. Consensus paper: towards a systems-level view of cerebellar function: the interplay between cerebellum, basal ganglia, and cortex. *The Cerebellum*, 16:203–229, 2017.

[13] Christopher H Chen, Rachel Fremont, Eduardo E Arteaga-Bracho, and Kamran Khodakhah. Short latency cerebellar modulation of the basal ganglia. *Nature neuroscience*, 17(12):1767–1775, 2014.

[14] Jane Y Chen, Elizabeth A Wang, Carlos Cepeda, and Michael S Levine. Dopamine imbalance in huntington's disease: a mechanism for the lack of behavioral flexibility. *Frontiers in neuroscience*, 7:49199, 2013.

[15] Ignasi Clavera, Violet Fu, and Pieter Abbeel. Model-augmented actor-critic: Backpropagating through paths. *arXiv preprint arXiv:2005.08068*, 2020.

[16] P Damier, EC Hirsch, Y Agid, and AM10430829 Graybiel. The substantia nigra of the human brain: Ii. patterns of loss of dopamine-containing neurons in parkinson's disease. *Brain*, 122 (8):1437–1448, 1999.

[17] Paul Dean and John Porrill. The importance of marr's three levels of analysis for understanding cerebellar function. *Computational Theories and their Implementation in the Brain: The legacy of David Marr*, 79, 2016.

[18] Alberto Del Arco and Francisco Mora. Prefrontal cortex–nucleus accumbens interaction: in vivo modulation by dopamine and glutamate in the prefrontal cortex. *Pharmacology Biochemistry and Behavior*, 90(2):226–235, 2008.

[19] Ashesh K Dhawale, Maurice A Smith, and Bence P Ölveczky. The role of variability in motor learning. *Annual review of neuroscience*, 40:479–498, 2017.

[20] Ashesh K Dhawale, Yohsuke R Miyamoto, Maurice A Smith, and Bence P Ölveczky. Adaptive regulation of motor variability. *Current Biology*, 29(21):3551–3562, 2019.

[21] Kenji Doya. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural networks*, 12(7-8):961–974, 1999.

[22] Kenji Doya. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current opinion in neurobiology*, 10(6):732–739, 2000.

[23] Richard P Dum and Peter L Strick. An unfolded map of the cerebellar dentate nucleus and its projections to the cerebral cortex. *Journal of neurophysiology*, 89(1):634–639, 2003.

[24] Stan B Floresco and Orsolya Magyar. Mesocortical dopamine modulation of executive functions: beyond working memory. *Psychopharmacology*, 188:567–585, 2006.

[25] Joseph M Galea, Alejandro Vazquez, Neel Pasricha, Jean-Jacques Orban de Xivry, and Pablo Celnik. Dissociating the roles of the cerebellum and motor cortex during adaptive learning: the motor cortex retains what the cerebellum learns. *Cerebral cortex*, 21(8):1761–1770, 2011.

[26] Joseph M Galea, Elizabeth Mallia, John Rothwell, and Jörn Diedrichsen. The dissociable effects of punishment and reward on motor learning. *Nature neuroscience*, 18(4):597–602, 2015.

[27] Michele Garibbo, Casimir Ludwig, Nathan Lepora, and Laurence Aitchison. What deep reinforcement learning tells us about human motor learning and vice-versa. *arXiv preprint arXiv:2208.10892*, 2022.

[28] Juan Manuel Gutierrez-Garralda, Pablo Moreno-Briseño, Marie-Catherine Boll, Consuelo Morgado-Valle, Aurelio Campos-Romo, Rosalinda Diaz, and Juan Fernandez-Ruiz. The effect of p arkinson's disease and h untington's disease on human visuomotor learning. *EJN*, 38(6): 2933–2940, 2013.

[29] Suzanne N Haber. Corticostriatal circuitry. *Dialogues in clinical neuroscience*, 18(1):7–21, 2016.

[30] Alkis M Hadjiosif, John W Krakauer, and Adrian M Haith. Did we get sensorimotor adaptation wrong? implicit adaptation as direct policy updating rather than forward-model-based learning. *Journal of Neuroscience*, 41(12):2747–2761, 2021.

[31] Masashi Hamada, Gionata Strigaro, Nagako Murase, Anna Sadnicka, Joseph M Galea, Mark J Edwards, and John C Rothwell. Cerebellar modulation of human associative plasticity. *The Journal of physiology*, 590(10):2365–2374, 2012.

[32] Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. *Advances in neural information processing systems*, 28, 2015.

[33] Gordon Holmes. The symptoms of acute cerebellar injuries due to gunshot injuries. *Brain*, 40 (4):461–535, 1917.

[34] J Hore, B Wild, and HC Diener. Cerebellar dysmetria at the elbow, wrist, and fingers. *Journal of neurophysiology*, 65(3):563–571, 1991.

[35] Eiji Hoshi, Léon Tremblay, Jean Féger, Peter L Carras, and Peter L Strick. The cerebellum communicates with the basal ganglia. *Nature neuroscience*, 8(11):1491–1493, 2005.

[36] Jonas A Hosp, Ana Pekanovic, Mengia S Rioult-Pedotti, and Andreas R Luft. Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning. *Journal of Neuroscience*, 31(7):2481–2487, 2011.

[37] James C Houk and James L Adams. 13 a model of how the basal ganglia generate and use neural signals that. *Models of information processing in the basal ganglia*, page 249, 1995.

[38] Court Hull. Prediction signals in the cerebellum: beyond supervised motor learning. *elife*, 9: e54073, 2020.

[39] Noritaka Ichinohe, Fumiaki Mori, and Kazuhiko Shoumura. A di-synaptic projection from the lateral cerebellar nucleus to the laterodorsal part of the striatum via the central lateral nucleus of the thalamus in the rat. *Brain research*, 880(1-2):191–197, 2000.

[40] Masao Ito. Neurophysiological aspects of the cerebellar motor control system. *Int. J. Neurol.*, 7:126–179, 1970.

[41] Jun Izawa and Reza Shadmehr. Learning from sensory and reward prediction errors during motor adaptation. *PLoS computational biology*, 7(3):e1002012, 2011.

[42] Jun Izawa, Sarah E Criscimagna-Hemminger, and Reza Shadmehr. Cerebellar contributions to reach adaptation and learning sensory consequences of action. *Journal of Neuroscience*, 32 (12):4230–4239, 2012.

[43] Daphna Joel, Yael Niv, and Eytan Ruppin. Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural networks*, 15(4-6):535–547, 2002.

[44] Michael I Jordan and David E Rumelhart. Forward models: Supervised learning with a distal teacher. In *Backpropagation*, pages 189–236. Psychology Press, 2013.

[45] Mitsuo Kawato, Kazunori Furukawa, and Ryoji Suzuki. A hierarchical neural-network model for control and learning of voluntary movement. *Biological cybernetics*, 57:169–185, 1987.

[46] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[47] Asha Kishore, Praveen James, Traian Popa, Arun Thejaus, Parvathy Rajeswari, Gangadhara Sarma, Syam Krishnan, and Sabine Meunier. Plastic responsiveness of motor cortex to paired associative stimulation depends on cerebellar input. *Clinical Neurophysiology*, 132(10): 2493–2502, 2021.

[48] Jeffrey A Kleim, Scott Barbay, and Randolph J Nudo. Functional reorganization of the rat motor cortex following motor skill learning. *Journal of neurophysiology*, 80(6):3321–3325, 1998.

[49] John W Krakauer. Motor learning and consolidation: the case of visuomotor rotation. *Progress in motor control: a multidisciplinary perspective*, pages 405–421, 2009.

[50] John W Krakauer, Alkis M Hadjiosif, Jing Xu, Aaron L Wong, and Adrian M Haith. Motor learning. *Compr Physiol*, 9(2):613–663, 2019.

[51] Łukasz Kuśmierz, Takuya Isomura, and Taro Toyoizumi. Learning with three factors: modulating hebbian plasticity with errors. *Current opinion in neurobiology*, 46:170–177, 2017.

[52] Timothy P Lillicrap, Pablo Moreno-Briseño, Rosalinda Diaz, Douglas B Tweed, Nikolaus F Troje, and Juan Fernandez-Ruiz. Adapting to inversion of the visual field: a new twist on an old problem. *Experimental brain research*, 228:327–339, 2013.

[53] Jack Lindsey and Ashok Litwin-Kumar. Action-modulated midbrain dopamine activity arises from distributed control policies. *NeurIPS*, 35:5535–5548, 2022.

[54] Andreas R Luft, Manuel M Buitrago, Thomas Ringer, Johannes Dichgans, and Jörg B Schulz. Motor skill learning depends on protein synthesis in motor cortex after training. *Journal of Neuroscience*, 24(29):6515–6520, 2004.

[55] Mario Manto. Mechanisms of human cerebellar dysmetria: experimental evidence and current conceptual bases. *Journal of neuroengineering and rehabilitation*, 6:1–18, 2009.

[56] David Marr. A theory of cerebellar cortex. *J Physiol*, 202:437–470, 1969.

[57] TA Martin, JG Keating, HP Goodkin, AJ Bastian, and WT Thach. Throwing while looking through prisms: I. focal olivocerebellar lesions impair adaptation. *Brain*, 119(4):1183–1198, 1996.

[58] Sarah Melzer, Mariana Gil, David E Koser, Magdalena Michael, Kee Wui Huang, and Hannah Monyer. Distinct corticostriatal gabaergic neurons modulate striatal output neurons and motor activity. *Cell reports*, 19(5):1045–1055, 2017.

[59] Frank A Middleton and Peter L Strick. Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain research reviews*, 31(2-3):236–250, 2000.

[60] Frank A Middleton and Peter L Strick. Cerebellar projections to the prefrontal cortex of the primate. *Journal of neuroscience*, 21(2):700–712, 2001.

[61] Demetrio Milardi, Angelo Quartarone, Alessia Bramanti, Giuseppe Anastasi, Salvatore Bertino, Gianpaolo Antonio Basile, Piero Buonasera, Giorgia Pilone, Giuseppe Celeste, Giuseppina Rizzo, et al. The cortico-basal ganglia-cerebellar network: past, present and future perspectives. *Frontiers in systems neuroscience*, 13:61, 2019.

[62] Katiuska Molina-Luna, Ana Pekanovic, Sebastian Röhrich, Benjamin Hertler, Maximilian Schubring-Giese, Mengia-Seraina Rioult-Pedotti, and Andreas R Luft. Dopamine in motor cortex is necessary for skill learning and synaptic plasticity. *PloS one*, 4(9):e7082, 2009.

[63] P Read Montague, Peter Dayan, and Terrence J Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, 16(5): 1936–1947, 1996.

[64] A Nieoullon, A Cheramy, and J Glowinski. Release of dopamine in both caudate nuclei and both substantia nigrae in response to unilateral stimulation of cerebellar nuclei in the cat. *Brain Research*, 148(1):143–152, 1978.

[65] André Nieoullon and Nicole Dusticier. Changes in dopamine release in caudate nuclei and substantia nigrae after electrical stimulation of the posterior interposate nucleus of cat cerebellum. *Neuroscience letters*, 17(1-2):167–172, 1980.

[66] Ali A Nikooyan and Alaa A Ahmed. Reward feedback accelerates motor learning. *Journal of neurophysiology*, 113(2):633–646, 2015.

[67] Yael Niv. Reinforcement learning in the brain. *Mathematical Psychology*, 53(3):139–154, 2009.

[68] Randoff J Nudo, Garrett W Milliken, W Merzenich Jenkins, and Michæl M Merzenich. Use-dependent alterations of movement representations in primary motor cortex of adult squirrel monkeys. *The Journal of neuroscience*, 16(2):785, 1996.

[69] Mark G Packard and Barbara J Knowlton. Learning and memory functions of the basal ganglia. *Annual review of neuroscience*, 25(1):563–593, 2002.

[70] Dimitrios J Palidis, Joshua GA Cashaback, and Paul L Gribble. Neural signatures of reward and sensory error feedback processing in motor learning. *Journal of neurophysiology*, 121(4): 1561–1574, 2019.

[71] Joseph Pemberton, Ellen Boven, Richard Apps, and Rui Ponte Costa. Cortico-cerebellar networks as decoupling neural interfaces. *Advances in neural information processing systems*, 34:7745–7759, 2021.

[72] Andrew J Peters, Julie MJ Fabre, Nicholas A Steinmetz, Kenneth D Harris, and Matteo Carandini. Striatal activity topographically reflects cortical activity. *Nature*, 591(7850): 420–425, 2021.

[73] Ludivine Pidoux, Pascale Le Blanc, Carole Levenes, and Arthur Leblois. A subcortical circuit linking the cerebellum to the basal ganglia engaged in vocal learning. *Elife*, 7:e32167, 2018.

[74] T Popa, B Velayudhan, C Hubsch, S Pradeep, E Roze, M Vidailhet, S Meunier, and A Kishore. Cerebellar processing of sensory inputs primes motor cortex plasticity. *Cerebral cortex*, 23(2): 305–314, 2013.

[75] Kasja Rabe, Ofer Livne, Elke R Gizewski, Volker Aurich, Andreas Beck, Dagmar Timmann, and Opher Donchin. Adaptation to visuomotor rotation and force field perturbation is correlated to different brain areas in patients with cerebellar degeneration. *Journal of neurophysiology*, 101(4):1961–1971, 2009.

[76] Sairam Ramesh and Arosh S Perera Molligoda Arachchige. Depletion of dopamine in parkinson's disease and relevant therapeutic options: A review of the literature. *AIMS neuroscience*, 10(3):200, 2023.

[77] John NJ Reynolds and Jeffery R Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural networks*, 15(4-6):507–521, 2002.

[78] Mengia-Seraina Rioult-Pedotti, Ana Pekanovic, Clement Osei Atiemo, John Marshall, and Andreas Rüdiger Luft. Dopamine promotes motor cortex plasticity and motor skill learning via plc activation. *PloS one*, 10(5):e0124986, 2015.

[79] Ankit Rohatgi. Webplotdigitizer: Version 4.5, 2021. URL `https://automeris.io/WebPlotDigitizer`.

[80] Jerome N Sanes, Bozhidar Dimitrov, and Mark Hallett. Motor learning in patients with cerebellar dysfunction. *Brain*, 113(1):103–120, 1990.

[81] Daniel Scherman, Claire Desnos, François Darchen, Pierre Pollak, France Javoy-Agid, and Yves Agid. Striatal dopamine deficiency in parkinson's disease: role of aging. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 26(4):551–557, 1989.

[82] Markus M Schugens, Caterina Breitenstein, Hermann Ackermann, and Irene Daum. Role of the striatum and the cerebellum in motor skill acquisition. *Behavioural neurology*, 11(3): 149–157, 1998.

[83] Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[84] Jeremy K Seamans and Charles R Yang. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in neurobiology*, 74(1):1–58, 2004.

[85] Roy V Sillitoe and Alexandra L Joyner. Morphology, molecular codes, and circuitry produce the three-dimensional complexity of the cerebellum. *Annu. Rev. Cell Dev. Biol.*, 23:549–577, 2007.

[86] Gordon D Smith. *Numerical solution of partial differential equations: finite difference methods*. Oxford university press, 1985.

[87] David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7):1025–1033, 2015.

[88] Amanda S Therrien, Daniel M Wolpert, and Amy J Bastian. Effective reinforcement learning following cerebellar damage requires a balance between exploration and motor noise. *Brain*, 139(1):101–114, 2016.

[89] Amanda S Therrien, Matthew A Statton, and Amy J Bastian. Reinforcement signaling can be used to reduce elements of cerebellar reaching ataxia. *The Cerebellum*, 20(1):62–73, 2021.

[90] Dmitrii I Todorov, Robert A Capps, William H Barnett, Elizaveta M Latash, Taegyo Kim, Khaldoun C Hamade, Sergey N Markin, Ilya A Rybak, and Yaroslav I Molkov. The interplay between cerebellum and basal ganglia in motor adaptation: A modeling study. *PLoS One*, 14 (4):e0214926, 2019.

[91] Jonathan S Tsay, Guy Avraham, Hyosub E Kim, Darius E Parvin, Zixuan Wang, and Richard B Ivry. The effect of visual uncertainty on implicit motor adaptation. *Journal of neurophysiology*, 125(1):12–22, 2021.

[92] Ya-weng Tseng, Jorn Diedrichsen, John W Krakauer, Reza Shadmehr, and Amy J Bastian. Sensory prediction errors drive cerebellum-dependent adaptation of reaching. *Journal of neurophysiology*, 98(1):54–62, 2007.

[93] Shinichiro Tsutsumi, Maya Yamazaki, Taisuke Miyazaki, Masahiko Watanabe, Kenji Sakimura, Masanobu Kano, and Kazuo Kitamura. Structure–function relationships between aldolase c/zebrin ii expression and complex spike synchrony in the cerebellum. *Journal of Neuroscience*, 35(2):843–852, 2015.

[94] Elinor Tzvi, Sebastian Loens, and Opher Donchin. Mini-review: the role of the cerebellum in visuomotor adaptation. *The Cerebellum*, 21(2):306–313, 2022.

[95] Katinka van der Kooij, Nina M van Mastrigt, Emily M Crowe, and Jeroen BJ Smeets. Learning a reach trajectory based on binary reward feedback. *Scientific Reports*, 11(1):2667, 2021.

[96] Tianhe Wang and Jordan A Taylor. Implicit adaptation to mirror reversal is in the correct coordinate system but the wrong direction. *Journal of neurophysiology*, 126(5):1478–1489, 2021.

[97] Samantha Washburn, Maritza Oñate, Junichi Yoshida, Jorge Vera, Ramakrishnan KB, Leila Khatami, Farzan Nadim, and Kamran Khodakhah. Cerebellum directly modulates the substantia nigra dopaminergic activity. *bioRxiv*, pages 2022–05, 2022.

[98] Thomas C Watson, Nadine Becker, Richard Apps, and Matthew W Jones. Back to front: cerebellar connections and interactions with the prefrontal cortex. *Frontiers in systems neuroscience*, 8:4, 2014.

[99] Kunlin Wei and Konrad Körding. Uncertainty of feedback and state estimation determines the speed of motor adaptation. *Frontiers in computational neuroscience*, 4:1151, 2010.

[100] Daniel M Wolpert, R Chris Miall, and Mitsuo Kawato. Internal models in the cerebellum. *Trends in cognitive sciences*, 2(9):338–347, 1998.

[101] Daniel M Wolpert, Jörn Diedrichsen, and J Randall Flanagan. Principles of sensorimotor learning. *Nature reviews neuroscience*, 12(12):739–751, 2011.

[102] Le Xiao, Caroline Bornmann, Laetitia Hatstatt-Burklé, and Peter Scheiffele. Regulation of striatal cells and goal-directed behavior by cerebellar outputs. *NatComms*, 9(1):3133, 2018.

[103] Junichi Yoshida, Maritza Oñate, Leila Khatami, Jorge Vera, Farzan Nadim, and Kamran Khodakhah. Cerebellar contributions to the basal ganglia influence motor coordination, reward processing, and movement vigor. *Journal of Neuroscience*, 42(45):8406–8415, 2022.

A Appendix

A.1 Line drawing task

Here, we want to show our action gradient model can support learning in a more complex task, involving a recurrent policy [e.g., stored in the motor cortex 87]. In this task, a recurrent policy [e.g., encoded in the motor cortex 87] needs to learn to control an arm to draw six target (straight) lines. Each line must be drawn over 10 time steps, t, given an initial cue provided at $t = 0$ only. The experimental details are further described in Appendix A.4.5. Fig. 4b shows that when the policy synaptic plasticity, $\Delta\phi$, is exclusively driven by the dopaminergic projections (i.e., conveying RBL action gradients) with 0% cortico-cerebellar contribution (i.e., conveying EBL action gradient), the resulting policy presents very high ataxia. Additionally, we also observe slower learning performance (Fig. 4a, in red). Conversely, as we increase the percentage of cortico-cerebellar contribution to the (policy) synaptic plasticity, the ataxia score decreases dramatically as well as achieve faster learning (Fig. 4a, in green and blue). Interestingly, we find the best policy, in terms of lowest ataxia score, is achieved when learning is primarily driven by the coritico-cerebellar signals (i.e., 75%) with some contribution by the dopaminergic projections (i.e., 25%). Although this finding requires further exploration, we speculate it may provide some insights on why the brain relies on close cerebellar and basal ganglia interactions during learning, instead of relying on either one exclusively [12].

Fig. 4c shows when learning is driven by cortico-cerebellar signals (i.e., EBL), task performance is not affected by the lack of continuous sensory feedback (e.g., sensory feedback is only available every 2 or 3 time steps). This is also the case when learning is driven by 'Mixed' cortico-cerebellar and dopamine signals (i.e., 'Mixed' condition, $\beta = 0.5$). Conversely, when learning only reflects dopamine-dependent RBL with no cerebellar contribution (DA-only), performance drastically worsens in the absence of continuous sensory feedback. This finding suggest cortico-cerebellar feedback may be crucial to maintain learning in the absence of continuous sensory feedback, even when learning is partially driven by dopamine-dependent feedback (i.e., 'Mixed' condition). To our knowledge, this prediction requires testing, although it aligns with the broad role of the cerebellum in facilitating learning [71, 8]. Finally, Fig. 4d-e show both dopamine-dependent (**d**) and cortico-cerebellar dependent (**e**) activity should decrease over the course of learning, reflecting task acquisition. Interestingly, dopamine-dependent activity seems to present higher variability relative to the cortico-cerebellar dependent activity, reflecting the higher variability of the RBL action-gradient relative to the EBL one. To our knowledge, this prediction requires testing.

A.2 Implausibility of RBL and EBL training separate policies

In this section, we want to investiagte the possibility that the brain possesses two separate policies for RBL and EBL (e.g., one stored in the cerebellum and another one stored in the basal ganglia), which are only combined at the action selection stage [e.g., 41, 90]. In particular, we assess two main ways in which the brain may combine separate RBL and EBL policies. In the first case, the output of the EBL and of the RBL policies could be summed to give the current action (i.e., $a = a^{\text{EBL}} + a^{\text{RBL}}$). In the second case, the two policy outputs could be combined through a weighted sum, enabling to prioritise one policy over the other when selecting the current action (i.e., $a = \alpha\, a^{\text{EBL}} + (1-\alpha)\, a^{\text{RBL}}$). For instance, in the presence of reward signals, but no sensory feedback, the brain could exclusively rely on the (most updated) RBL policy (i.e., $\alpha = 0$). Conversely, in the presence of optimal sensory feedback, the brain could prioritise the EBL policy (i.e., $\alpha = 1$). Note, this weighted-sum case is in principle similar to what we proposed in eq. 4, but where the weighted sum occurs at the level of actions (based on two separate polices) instead of at level of teaching signals (i.e., action gradient). We test these two-policy set-ups in a targeted arm reaching task, where the agent needs to learn to reach for a target location under optimal conditions (i.e., minimal sensory noise and no perturbations). This task is based on standard sensorimotor adaptation paradigms used to probe motor adaptation in humans [e.g., 41, 30, 49]. It requires the agent to learn to output the correct reaching angles to reach a series of targets across several trials. The targets are placed at different angles from the agent, thus requiring the agent to learn a different reaching angle for each target (see Appendix A.4.1 for further details on the model).

Fig. 5a shows that combining the two policies through a weighted sum (i.e., $\alpha -$ weighting) leads to better performance than performing a simple sum of the two polices' outputs (i.e., No $\alpha -$ weighting). This is because for the given task with optimal sensory feedback, EBL seems to be the optimal

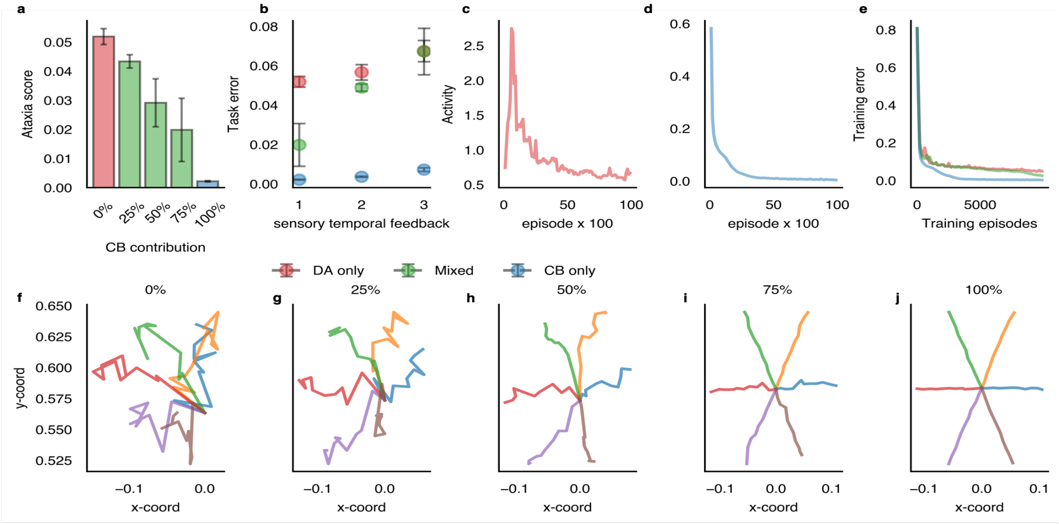

Figure 4: ***a-b**) Training accuracy (**a**) and ataxia score (**b**) across different percentages of cortico-cerebellar (CC) contribution to learning on the line drawing task. **c**) Task performance for different temporal feedback across dopamine-dependent learning (DA), cortico-cerebellar learning (CC) and a mixture of the two (Mixed). **d-e**) Dopamine-dependent (**d**) and cortico-cerebellar (**e**) activity across learning (i.e., in terms of the norm of the respective RBL and EBL action gradients). **f-j**) Resulting line drawings across different percentages of cortico-cerebellar (CC) contribution to learning. DA: dopamine-dependent learning, CC: cortico-cerebellar learning, Mixed: dopamine and cortico-cerebellar dependent (i.e., $\beta = 0.5$).*

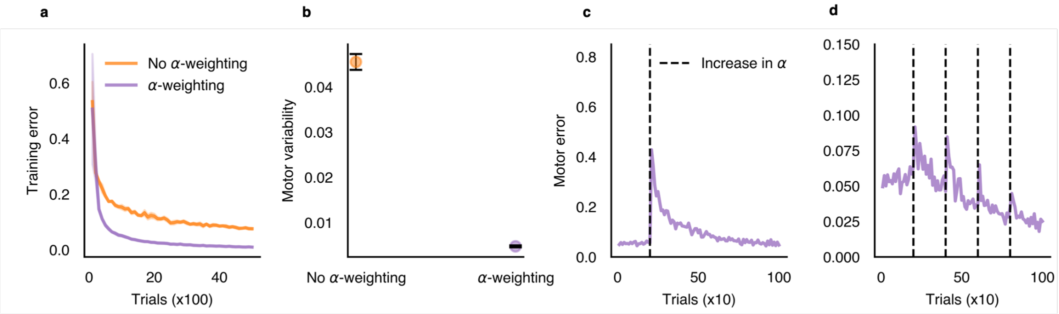

Figure 5: *Learning performance for separate EBL and RBL policies, where the two policies are combined by either a simple sum (No $\alpha -$ weighting) or based on a weighted sum ($\alpha -$ weighting). **a**) the weighted-sum approach achieves better learning performance than the simple sum approach. **b**) the weighted-sum approach achieves lower motor variability than the simple sum approach. **c-d**) the weighted-sum approach suffers from a drop in performance whenever there is a gradual (**c**) or sudden (**d**) change in sensory vs reward feedback.*

learning strategy and, the weighted-sum approach ($\alpha -$ weighting) is able to suppress the contribution of the (sub-optimal) RBL policy (i.e., by setting $\alpha = 1$). The RBL policy cannot be suppressed in the simple-sum approach (No $\alpha -$ weighting), negatively affecting the performance. Furthermore, Fig. 5b shows that combining the two policies through a simple sum leads to higher motor variability than performing the weighted sum, due to the impossibility of the simple-sum approach to suppress the extra motor variability induced by the RBL policy. Indeed, we later show that without assuming some weighted combination of RBL and EBL contributions, it is not possible to explain the experimental evidence that (human) motor variability changes depending on the quality of the sensory feedback [41]. Therefore, if the brain holds two separate policies for RBL and EBL, then the brain likely rely on a weighted sum of the two polices (i.e., $\alpha -$ weighting), instead of a simple sum. However, we now show this two-policy weighted-sum approach ($\alpha -$ weighting) is also unlikely to take place in the brain. To show this, we assume a reaching setting where only the rewards are initially available with no sensory feedback (i.e., $\alpha = 0$). After some initial trials, we either abruptly (Fig. 5c) or gradually (Fig. 5d) introduce sensory feedback (i.e., marked by an increase in α), thus enabling the EBL policy to contribute to the actions. Crucially, Fig. 5c-d shows that whenever there is a change in the sensory feedback (i.e., altering the contribution of each policy), the performance suddenly worsens. This occurs because the two policies are no longer synchronized. Crucially, the brain is exposed to constant changes to the quality of sensory feedback in the external world. Additionally, performance should likely improve if better sensory feedback becomes available. Therefore, we believe it is implausible the brain holds two separate policies, one trained with EBL and another trained by RBL. We think it is more plausible the brain combines EBL and RBL at the level of teaching signals instead of at the level of actions (i.e., as in eq. 4).

A.3 Policy representations

As we briefly described in section 2, the actions (motor commands) are assumed to follow a Gaussian policy,

$$a \sim \pi_\phi(a_t \mid s_t) = \mathcal{N}(a_t \mid \mu_\phi(s_t), \Sigma_\phi(s_t) + \Sigma_{\text{fixed}}) \tag{5}$$

where t denotes the current trial, the mean, μ_ϕ, and (diagonal) covariance/variance, Σ_ϕ, are parametrized by synaptic weights, ϕ (e.g., the synaptic weights of the area encoding the policy). We also assume a certain degree of motor variability to be fixed and non reducible (i.e., Σ^{fixed}), since some degree of human motor variability seems always present [19]. These (policy) synaptic weights, ϕ, are learned based on the action-gradient model introduced in section 3. In the sections 5.1, A.2 and the visuomotor reversal task A.4.2, the policy parametrization consists of a linear model, taking the desired target as input (i.e., $s = y^*$) and, outputting the policy mean, μ_ϕ, and variance, σ^2_ϕ (i.e., for the 1-dimensional actions),

$$\mu_\phi = y^* \phi_1 + \phi_2 \tag{6}$$

$$\sigma^2_\phi = y^* \phi_3 + \phi_4 \tag{7}$$

Nore, In in the experiments with separate policies for EBL and RBL (Appendix A.2), we use two separate (linear) policies, where the RBL policy is exclusively updated based on RBL (section 2), while the EBL policy is exclusively updated based on EBL (section 2) [e.g., 43, 41, 30, 27]. Note, this (linear) policy set-up is standard practice in computational models of motor adaptation, where the action, a, typically represents the arm reaching angle [e.g., 41, 30, 27].

In all the motor learning tasks involving the (non-linear) two-joint arm model, the (Gaussian) policy is parametrized by 1-hidden layer feedforward neural network, predicting the policy mean, μ_ϕ and covariance Σ_ϕ, by taking a cue as input (i.e., the cue encoding different targets in the tasks). In this setting, the actions sampled from the policy were two dimensional, consisting of the two join angles controlling the arm position in xy-coordinates (see Appendix A.5). Finally, in the (non-linear) line drawing task (Appendix A.1), the (Gaussian) policy parametrization consists of a recurrent neural network. This is because the policy has to maintain a memory of the initial cue, which is only presented at the start of the task. We speculate this (recurrent) policy may be encoded in the motor cortex, which has previously been associated to recurrent neural networks [87].

A.4 Motor learning tasks

The objective of all motor learning tasks is to find the policy parameters, ϕ, that maximises the reward, r. For each trial t, the scalar reward is computed as the (negative) value of the squared distance between the sensory outcome, y_t, (e.g., the arm position) and the location of the target, denoted as y^*,

$$r_t = -(y^* - y_t)^2 \tag{8}$$

So the maximum reward is achieved at the target location. Both the RBL and EBL action gradients attempt to maximise this reward signal (see section 3).

A.4.1 Modelling Izawa and Shadmer arm reaching tasks

We employ 2 different targeted arm reaching tasks to reproduce Izawa and Shadmehr [41]'s results on motor variability and motor generalisation under RBL and EBL. In line with previous models of reaching adaptation [e.g., 41, 30, 27], we model the reaching dynamics based on a linear motor model, where the terminal arm position, y, is related to the action, a (i.e., reaching angle) as following,

$$y_t = wa_t + c + \epsilon^{\text{sensory}} \tag{9}$$

where $\epsilon^{\text{sensory}} \sim \mathcal{N}(0, \sigma^{\text{sensory}})$ controls the amount of sensory noise over the terminal arm position In section 5.1, the task follows the exact same set-up as Izawa and Shadmehr [41], where the model had to reach for a target location, y^*, under a visuomotor rotation, which was gradually increased over 320 reaching trials (i.e., 1-degree increase every 40 trials up to 8 degrees of rotation perturbation) [see 41, for more details]. To do so, we pre-train a policy to reach under normal conditions. Note, this step is not necessary with human participants, since they already know how to reach under normal conditions. Next, we assess the policy ability to adapt to the rotation perturbation under RBL, EBL and a mixture of the two (as described in section 3). In the motor generalisation task, we also followed the exact same set-up as Izawa and Shadmehr [41], where the policy generalisation performance was tested on targeted located at $[-30, -20, -10, 0, 10, 20, 30]$ degrees.

A.4.2 Modelling the sensorimotor rotation and reversal reaching tasks

Both tasks involved three phases, 1) an initial baseline phase with no perturbation, 2) a perturbation phase (i.e., where the rotation or reversal pertubation was applied), 3) a washout phase, where the perturbation is (suddenly)removed. Before the baseline phase, we pre-trained a policy to perform the task under normal conditions (i.e., without the perturbation), so that we could get the baseline performance Note, this step is not necessary with human participants, since they already know how to perform the task under normal conditions (e.g., they already know how to reach). For the sensorimotor rotation task, we used the exact same task set-up as Tseng et al. [92], where we modelled the targeted arm reaches based on the two-joint kinematic arm model described in Appendix A.5. Briefly, the agent had to reach for 3 different target location placed at angle degrees of [-45, 0, 45] from the initial arm position. During the perturbation phase, the visual outcomes in xy-coordinates were rotated of 20 degree relative to its true position and the agent had to learn to adapt its reaches to correct for the rotation. The residual error was calculated as the averaged reaching error (i.e., in terms of the euclidean distance from the target) of the last 25 trials of the adaptation phase, reflecting the same error measure used by Tseng et al. [92] for the experimental data. Further experimental details can be found at Tseng et al. [92]. For the sensorimotor reversal, we used the exact same task set-up as Gutierrez-Garralda et al. [28], where we model dart throwing based on a simple linear model similar the one described in Appendix A.4.1. The actions denoted the angle direction at which the dart was thrown. For the reversal phase, we introduce a reversing of the perceived angle position of the target by 11.31 degrees, replicating Gutierrez-Garralda et al. [28]. The adaptation measure was defined as the mean difference between the final throw and the first throw of the perturbation condition, reflecting the same adaptation measure used by Gutierrez-Garralda et al. [28] for the experimental data. For the model predictions in Fig. 3d, the residual error was computed as the (angle) distance from the target angle for the last trial of the adaptation phase (i.e., how much the agent still had to adapt to successfully solve the task). Further experimental details can be found at Gutierrez-Garralda et al. [28].

A.4.3 Dysfunctional cerebellar predictions during reaching

For this investigation, we again modelled arm reaches based on the two-joint kinematic arm model described in Appendix A.5. However, we did not apply any perturbation to the arm reaches, the agent simply had to reach for a target under normal conditions. To do so, we pre-trained a policy so that the agent knew how to reach for the target (i.e., cerebellar patients usually know how to reach for targets before they experience a cerebellar lesion). Next, we perturbed the cerbellar predictions of the action-to-outcome gradient and measured what happened to the reaching endpoints as the agent kept reaching for the target (i.e., adapting its reaches based on small irreducible error). This set-up aims to replicate cerebellar patients that after their cerebellar lesion start performing motor tasks again. The perturbation consisted on switching the sign to one (or more) action-to-outcome gradient components predicted by the cerebellum. Based on the two-joint kinematic arm model, there are four action-to-outcome gradient components that could be perturbed. This is because the arm takes two dimensional actions as inputs (i.e., the two joint angles) and output a two dimensional vector, encoding the xy-coordinate of the arm position in space. We reported below the four action-to-outcome gradient components (i.e., the Jacobian matrix of the two-joint arm model),

$$\begin{pmatrix} \frac{dy_1}{da_1} & \frac{dy_1}{da_2} \\ \frac{dy_2}{da_1} & \frac{dy_2}{da_2} \end{pmatrix} \tag{10}$$

where, (y_1, y_2) denote the xy-coordinate of the arm position after taking the action (a_1, a_2), encoding the two joint angles (i.e., denoted as (ψ^1, ψ^2) in the arm model). Therefore, the cerebellar module learned to predict these four components in eq. 10. First, we induced a cerebellar perturbation by switching the sign to one component at the time, aiming to reproduce a focal cerebellar lesion. This resulted in the classical cerebellar dysmetria symptoms plotted in Fig. 3f,i,j, where each plot represented a different component that was perturbed (note, we found one component did not affect performance, that's why there are only 3 plots with perturbations). For Fig. 3g, we randomly switched the sign of a different cerebellar predicted component before each trial, aiming to reproduce a broader cerebellar lesion. The dysmetria score was computed as the amount of vertical displacement plus the amount of horizontal displacement from the target (i.e., in terms of euclidean distance), in line with classical measure of dysmetria [see 33].

A.4.4 Toy task

This task aimed to reproduce the complex non-linear and high-dimensional relation between actions, a, and sensory outcomes, y, in the 'real' world. To do so, we model the task dynamics with a high-dimensional non-linear feedforward neural network with fixed weights (i.e., fixed task dynamics). This network comprised a single hidden layer of 256 hidden units, taking 10-dimensional actions as input, a, and outputting a 3 dimensional outcome, y (e.g., xyz-coordiante in space). The goal of the task was the same as the previous tasks, learning a policy, π_ϕ, that achieve the desired outcome, y^*. Specifically, the task comprised of 10 targets, y^*, which the policy had to achieve. The policy consisted of a single hidden layer feedforward neural network (56 hidden units), taking a cue as input and outputing a Gaussian distribution over the actions (i.e., a Gaussian polciy). The cue served to distinguish the 10 different targets. Similarly to all the other tasks, the reward function, r, was the (negative) of the squared euclidean distance between the current output and the target output, (i.e., $r = -(y^* - y)^2$). The sensory noise levels were modelled by adding zero mean Gaussian noise of increasing standard deviations to the outcomes, y. The optimal β was determined by finding the β value that lead to learning the policy achieving the highest task reward across an entire training run with a fixed level of sensory noise (i.e., for 5000 trials/episodes).

A.4.5 Line drawing task

In the line drawing task, a recurrent (Gaussian) policy has to learn to control a two-joint arm model to draw six (target) straight lines, each over 10 time steps, t. An initial cue indicates which of the six (target) lines the policy has to draw at any given trial. Crucially, this cue is only provided at the initial time step, $t = 0$. Therefore, the policy has to recall the correct cue over the following 9 steps in order to draw the correct line. In this task, the reward is computed as following,

$$r = -\sum_{t \in \text{FB}} (x_t - x_t^*(s))^2 + (y_t - y_t^*(s))^2 \tag{11}$$

where $(x_t^*(s), y_t^*(s))$ denotes the target location at time t to draw the corresponding target straight line given the initial cue, s, while FB indicates the set of time steps at which a feedback is provided (i.e., controlling the sensory temporal feedback). In order to compute the 'directed' sensory error component, $\frac{dr}{dy}$, of the EBL action gradient, we assume the brain can directly differentiate the reward (function) relative to y. We think this is plausible since, in this task, the reward is again a simple (squared) distance in visual space. Alternatively, a (differentiable) reward function may be learned from practice. Conversely, we assume the cerebellar module learns to predict the action-to-outcome gradient component, $\frac{dy}{da}$, of the EBL action gradient (see section A.6). Finally, the ataxia score is computed as the total deviation between the arm position across all the 10 time steps and the required locations to draw the correct straight line given the cue,

$$\text{Ataxia score: } \sum_{t=0}^{10} \sqrt{(x_t - x_t^*(s))^2 + (y_t - y_t^*(s))^2} \tag{12}$$

This is in line with previous work [e.g., 8, 71].

A.5 Two-joint arm model

To model the arm reaches, we employ a 2-dimensional (non-linear) kinematic arm model [17],

$$x_t = l^1\ cos(\psi_t^1)\ +\ l^2\ cos(\psi_t^1 + \psi_t^2) \tag{13}$$

$$y_t = l^1\ sin(\psi_t^1)\ +\ l^2\ sin(\psi_t^1 + \psi_t^2) \tag{14}$$

where (x_t, y_t) denote the arm position in xy-coordinates at the trial t (i.e., modelling planar reaching movements). The variables ψ_t^1 and ψ_t^2 respectively represent the angles of the shoulder and elbow joints at trial t, while l^1 and l^2 respectively denote the (constant) upper and lower arm length. Therefore, at any given trial t, the policy has to provide the two corresponding joints' angles to control the position of the arm in space, $(\psi_t^1, \psi_t^2) \sim \pi_\phi(a_t \mid s)$.

A.6 Cerebellar module

In the system-level circuits (section 4), we assume the cerebellum learns to predict the action-to-outcome gradient component, $\frac{dy}{da}$, of the EBL action gradient [e.g., 21]. In line with previous work [71, 8], we represent the cerebellum as a one hidden-layer feedforward neural network. This cerebellar network is trained to predict the action-to-outcome gradient, $\frac{dy}{da}$, given the current action, a, (motor command) and the corresponding (sensory) outcome, y at the end of each trial. This predicted action-to-outcome gradient, $\frac{d\hat{y}}{da}$, is successively combined with the corresponding 'directed' sensory error component, $\frac{d\hat{r}}{dy}$, (i.e., coming from sensory cortical areas) to compute the EBL action gradient. In our experiments, we assume the cerebellar network is provided with the target action-to-outcome gradients, $\frac{dy}{da}$. Hence, we could train the cerebellar network using a mean-squared loss between the network predictions and the target action-to-outcome gradients, for each task,

$$\mathcal{L}_{\text{cerebellum}} = \frac{1}{T}\sum_{t=0}^{T}\left(\frac{dy_t}{da_t} - f_{\text{cerebellum}}(a_t, y_t)\right)^2 \tag{15}$$

for T number of trials and $f_{\text{cerebellum}}(a_t, y_t)$ representing the cerebellar network. We followed this procedure for simplicity, since the scope of the paper was not explain how the cerebellum may learn the action-to-outcome gradients.

Nevertheless, in Fig. 1, we speculate in the brain, the cerebellum may learn these quantity based on finite-difference methods of gradient estimation [see 86]. Specifically, the cortex could keep track of small changes in (sensory) outcomes, Δy, relative to small changes to motor commands (actions) Δa. The cortex could send this information to the inferior olive, which in turn could compute a finite difference estimate of the action-to-outcome gradient (i.e., by simply computing a ratio, $\frac{\Delta y}{\Delta a}$), providing this as a target to the cerebellum [i.e., in line with well-known mechanism of the inferior olive providing a teaching signal to the cerbellum 100, 56, 40, 3].

A.7 Action-gradient update implementation

In order to compute the 'directed' sensory error component, $\frac{dr}{dy}$, of the EBL action gradient, we directly differentiate the reward (function) in eq. 8 relative to the observed sensory outcome y, to obtain an estimate $\frac{d\hat{r}}{dy}$. We think this is plausible in the brain since, in the tested tasks, the reward is a simple (squared) distance between the target, y^*, and the movement outcome, y, in visual space (i.e., eq. 8), suggesting the brain may directly estimate $\frac{d\hat{r}}{dy}$ from visual information. Alternatively, the brain may learn a (differentiable) reward function for more complex rewards. Conversely, we assume the action-to-outcome gradient component, $\frac{dy}{da}$, of the EBL action gradient to be estimated by a cerebellar module (see Appendix A.6). For instance, in the RL literature, the terms $\frac{d\hat{r}}{dy}$ and $\frac{d\hat{y}}{da}$ are typically estimated by differentiating through a learned model of the transitions, $(\hat{r}_t, \hat{y}_t) = f_w(s_t, a_t)$ for an initial state, s [i.e., model-based policy gradient 32]. Finally, to correctly estimate the product between the Mixed action gradient in eq 4 and the term driving synaptic plasticity, $\frac{d\mu_\phi}{d\phi}$, we employ the reparameterization trick. We need this because the policy is Gaussian implying the relation between actions, a, and the policy parameter, μ_ϕ, is stochastic [46].

We also report here the 'Mixed' action-gradient update for the variance parameter of the Gaussian policy, σ_ϕ^2, (i.e., reported here for 1-dimensional case for simplicity),

$$\frac{dr}{d\sigma_\phi^2} = \mathrm{E}_a \left(\overbrace{\frac{d\hat{r}}{dy}\frac{d\hat{y}}{da}}^{\text{EBL action gradient}} \beta + \overbrace{\frac{\delta((a-\mu_\phi)^2 - \sigma_\phi^2)}{\sigma_\phi^3}}^{\text{RBL action gradient}} (1-\beta) \right) \tag{16}$$

Note, the RBL action gradient on average increases the (policy) variance parameter whenever the RPE is negative (i.e., $\delta < 0$), while decreasing it whenever the RPE is positive (i.e., $\delta > 0$). Hence, under RBL, we should expect higher motor variability whenever we do not encounter rewards (i.e., leading to negative RPEs) and vice-versa.

A.8 Code repository

All the code is available at `https://anonymous.4open.science/r/ActionGradients-901B/`

A.9 Hyperparameters

Several hyperparameters are reported throughout the Appendix, all the rest can be found at `https://anonymous.4open.science/r/ActionGradients-901B/`.