

# Reinforcement Learning of Targeted Movement in a Spiking Neuronal Model of Motor Cortex

George L. Chadderdon<sup>1\*</sup>, Samuel A. Neymotin<sup>1,2</sup>, Cliff C. Kerr<sup>1,3</sup>, William W. Lytton<sup>1,4,5</sup>

**1** Department of Physiology and Pharmacology, State University New York Downstate, Brooklyn, New York, United States of America, **2** Department of Neurobiology, Yale University School of Medicine, New Haven, Connecticut, United States of America, **3** School of Physics, University of Sydney, Sydney, Australia, **4** Department of Neurology, State University New York Downstate, Brooklyn, New York, United States of America, **5** Kings County Hospital, Brooklyn, New York, United States of America

## Abstract

Sensorimotor control has traditionally been considered from a control theory perspective, without relation to neurobiology. In contrast, here we utilized a spiking-neuron model of motor cortex and trained it to perform a simple movement task, which consisted of rotating a single-joint “forearm” to a target. Learning was based on a reinforcement mechanism analogous to that of the dopamine system. This provided a global reward or punishment signal in response to decreasing or increasing distance from hand to target, respectively. Output was partially driven by Poisson motor babbling, creating stochastic movements that could then be shaped by learning. The virtual forearm consisted of a single segment rotated around an elbow joint, controlled by flexor and extensor muscles. The model consisted of 144 excitatory and 64 inhibitory event-based neurons, each with AMPA, NMDA, and GABA synapses. Proprioceptive cell input to this model encoded the 2 muscle lengths. Plasticity was only enabled in feedforward connections between input and output excitatory units, using spike-timing-dependent eligibility traces for synaptic credit or blame assignment. Learning resulted from a global 3-valued signal: reward (+1), no learning (0), or punishment (−1), corresponding to phasic increases, lack of change, or phasic decreases of dopaminergic cell firing, respectively. Successful learning only occurred when both reward and punishment were enabled. In this case, 5 target angles were learned successfully within 180 s of simulation time, with a median error of 8 degrees. Motor babbling allowed exploratory learning, but decreased the stability of the learned behavior, since the hand continued moving after reaching the target. Our model demonstrated that a global reinforcement signal, coupled with eligibility traces for synaptic plasticity, can train a spiking sensorimotor network to perform goal-directed motor behavior.

**Citation:** Chadderdon GL, Neymotin SA, Kerr CC, Lytton WW (2012) Reinforcement Learning of Targeted Movement in a Spiking Neuronal Model of Motor Cortex. PLoS ONE 7(10): e47251. doi:10.1371/journal.pone.0047251

**Editor:** Gennady Cymbalyuk, Georgia State University, United States of America

**Received:** June 15, 2012; **Accepted:** September 12, 2012; **Published:** October 19, 2012

**Copyright:** © 2012 Kerr et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Research supported by Defense Advanced Research Projects Agency (DARPA) grant N66001-10-C-2008 (<http://www.darpa.mil/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: georgec@neurosim.downstate.edu

## Introduction

Sensorimotor mappings, for example between proprioceptive input and motor output, are the basis for directed behavior, including foraging, locomotion, and object manipulation. Artificial neural networks, generally based on continuous unit states, have used a variety of learning algorithms to learn these mappings; examples include backpropagation [1,2], self-organizing maps [3], and temporal difference learning [4].

Artificial neural network models, as well as lumped control theory models, use processing units with continuous outputs which encode continuous rates or probabilities of firing. By contrast, recent models have begun to look more closely at biomimetic mechanisms by using spiking models for dynamics and spike-timing-dependent plasticity (STDP) for learning [5–12]. Spiking units offer the advantage of allowing us to explore multiple methods of neural encoding that are absent from continuous unit models. These include exposing possible roles of synchrony in perceptual feature binding and response selection [13], wave-front encoding [14,15], and other time-based codes. Physiologically, the degree of input spike synchrony is a major determinant of motor neuron activation [16].

Sensorimotor mappings can be thought of as stimulus-response mappings, suggesting reinforcement learning (RL) as a mechanism for learning. The essence of this learning mechanism was summarized over 100 years ago in Thorndike’s Law of Effect: stimulus-response mappings are strengthened by global reward and weakened by global punishment [17]. RL methods [18], including temporal-difference learning [4], have been used extensively in machine learning and offer an advantage over teacher-supervised learning methods in that they do not require a known desired output representation to match against the model’s current (behavioral) output. However, unlike unsupervised learning methods, they do offer some feedback regarding fitness of the behavior. A further framework for explaining motor RL is the perception-action-reward cycle [19]. The learning system is divided into an *actor*, mapping perceptions to actions (P to A), and a *critic* providing reward and punishment feedback to the actor [8,20,21]. To utilize this scheme, the naive actor must produce some actions. This is the role of *motor babble* [20,22,23], produced in our model via noise.

One challenge in the learning of actor/critic RL systems is the *distal reward or credit assignment problem* [7]: reinforcers are delivered after the behavior is complete, after synaptic and neuronal activations leading up to the output are no longer active. A

synaptic eligibility trace can be used to solve this problem: synapses are tagged to receive a credit or blame signal that arrives later [8]. Synapses tagged with eligibility traces, possibly mediated by transient phosphorylations [24] or dendritic  $\text{Ca}^{2+}$  currents [25,26], may be reinforced by the global reinforcement signals mediated by phasic reward bursts [27,28] and punisher dips [29] of dopamine cell firing from ventral tegmental area (VTA) projecting to cortical areas [30–34].

In this paper, we simulated a potential mechanism for the learning of sensorimotor mappings, using a biologically-inspired computational model consisting of spiking neuronal units whose synaptic weights are trained via global reward and punisher signals. This architecture was able to perform a stationary targeting task as long as both reward and punishment signals were present during the learning. Stable proprioceptive-to-motor command mappings mediated performance of the task.

## Methods

### Neuron Model

Individual neurons were modeled as event-driven, rule-based dynamical units with many of the key features found in real neurons, including adaptation, bursting, depolarization blockade, and voltage-sensitive NMDA conductance [35–40]. Event-driven processing provides a faster alternative to network integration: a presynaptic spike is an event that arrives after a delay at postsynaptic cells; this arrival is then a subsequent event that triggers further processing in the postsynaptic cells. Cells were parameterized as excitatory (E), fast-spiking inhibitory (I), and low-threshold-spiking inhibitory (IL; Table 1). Each cell had a membrane voltage state variable ( $V_m$ ), with a baseline value determined by a resting membrane potential parameter ( $V_{RMP}$ ). After synaptic input events, if  $V_m$  crossed spiking threshold ( $V_{th}$ ), the cell would fire an action potential and enter an absolute refractory period, lasting  $\tau_{AR}$ . After an action potential, an after-hyperpolarization voltage state variable ( $V_{AHP}$ ) was increased by a fixed amount  $W_{AHP}$  and then  $V_{AHP}$  was subtracted from  $V_m$ . Then  $V_{AHP}$  decayed exponentially (with time constant  $\tau_{AHP}$ ) to 0. To simulate voltage blockade, a cell could not fire if  $V_m$  surpassed the blockade voltage ( $V_{block}$ ). Relative refractory period was simulated after an action potential by increasing the firing threshold  $V_{th}$  by  $W_{RR}(V_{block} - V_{th})$ , where  $W_{RR}$  was a unitless weight parameter.  $V_{th}$  then decayed exponentially to its baseline value with time-constant  $\tau_{RR}$ .

In addition to the intrinsic membrane voltage state variable, each cell had four additional voltage state variables  $V_s$  corresponding to synaptic input. These represent AMPA, NMDA, and somatic and dendritic  $\text{GABA}_A$  synapses. At the times of input events, synaptic weights were updated by step-wise changes in  $V_s$ , which were then added to the cell's overall membrane voltage  $V_m$ . To allow for dependence on  $V_m$ , synaptic inputs changed  $V_s$  by  $dV = W_s(1 - V_m/E_s)$ , where  $W_s$  is the synaptic weight and  $E_s$  is

the reversal potential relative to  $V_{RMP}$ . The following values were used for the reversal potential  $E_s$ : AMPA, 65 mV; NMDA, 90 mV; and  $\text{GABA}_A$ , -15 mV. After synaptic input events, the synapse voltages  $V_s$  decay exponentially toward 0 with time constants  $\tau_s$ . The following values were used for  $\tau_s$ : AMPA, 20 ms; NMDA, 300 ms; somatic  $\text{GABA}_A$ , 10 ms; and dendritic  $\text{GABA}_A$ , 20 ms. The delays between inputs to dendritic synapses (AMPA, NMDA, dendritic  $\text{GABA}_A$ ) and their effects on somatic voltage were selected from a uniform distribution ranging between 3–5 ms, while the delays for somatic synapses (somatic  $\text{GABA}_A$ ) were selected from a uniform distribution ranging from 1.8–2.2 ms. Synaptic weights were fixed between a given set of populations except for those involved in learning (described below).

### System Design

The network system, shown in Fig. 1, consisted of (1) a simple one-joint “forearm,” with flexor and extensor muscles; (2) proprioceptive neurons, each tuned to fire at a specific joint angle; (3) sensory cells, which received spiking input from the proprioceptive cells; (4) motor command cells, which received spiking input from sensory cells and sent elbow rotation commands to the muscles; and (5) a reinforcement learning critic, which evaluated the change of hand-to-target visual error and sent a global reward or punisher training signal to the plastic synapses. The proprioceptive, sensory, and motor neurons were implemented using the model described above; further details on the system are provided below.

Input to the sensory cells was provided by 48 proprioceptive (P) cells, representing muscle lengths in 2 groups (flexor- and extensor-associated). Each was tuned to produce bursting approaching 100 Hz over a narrow range of adjacent, non-overlapping lengths. The cortical network consisted of both sensory and motor cell populations. The sensory (S) population included 96 excitatory sensory cells (ES cells), 22 fast spiking sensory interneurons (IS), and 10 low-threshold spiking sensory interneurons (ILS); similarly, the motor (M) network had 48 EM, 22 IM, and 10 ILM cells. The EM population was divided into two 24-cell subpopulations dedicated to extension and flexion, which projected to the extensor and flexor muscles, respectively.

Cells were connected probabilistically with connection densities and initial synaptic weights varying depending on pre- and postsynaptic cell types (Table 2). In addition to spikes generated by cells in the model, subthreshold Poisson-distributed spike inputs to each synapse of all units except the P and ES units were used to provide ongoing activity and babble (Table 3).

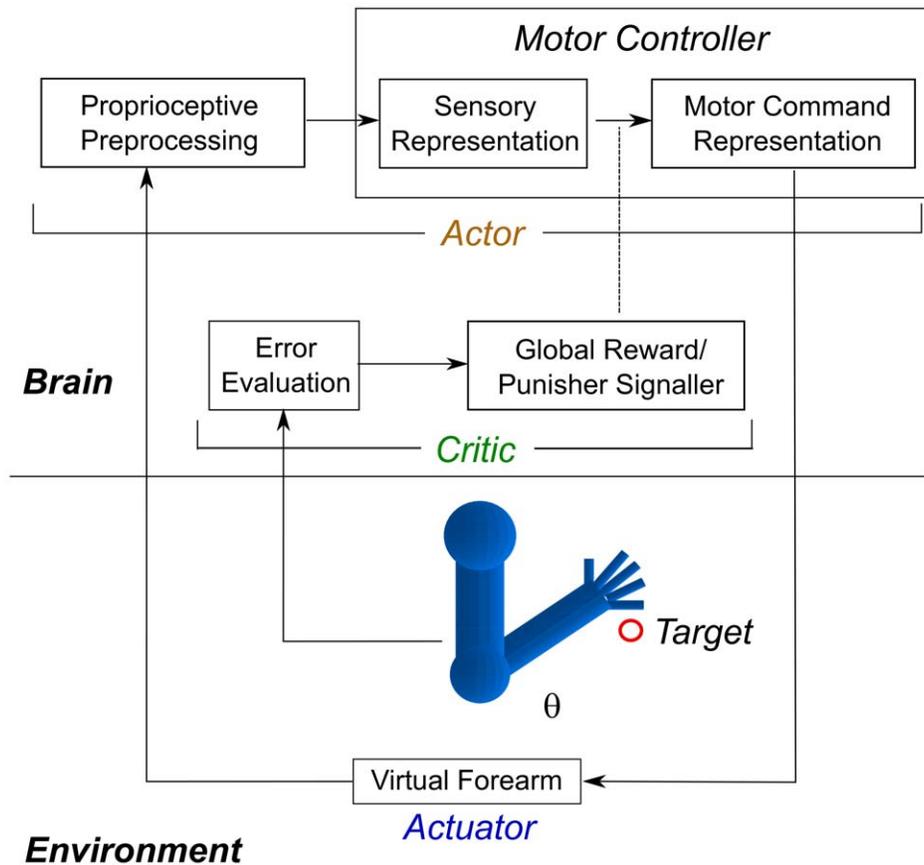
The virtual forearm consisted of a single segment of length  $l$  with a joint angle  $\theta$  that was allowed to vary from  $0^\circ$  (arm straight) to  $135^\circ$  (fully flexed). An extensor and flexor muscle (lengths  $m_{ext}$  and  $m_{flex}$ ) always reflected the current joint angle according to the following relationship:

**Table 1.** Parameters of the neuron model for each major population type.

Cell type	$V_{RMP}$ (mV)	$V_{th}$ (mV)	$V_{block}$ (mV)	$\tau_{AR}$ (ms)	$W_{RR}$	$\tau_{RR}$ (ms)	$W_{AHP}$ (mV)	$\tau_{AHP}$ (ms)
Excitatory	-65	-40	-25	5	0.75	8.0	1.0	400
Inhibitory	-63	-40	-10	2.5	0.25	1.5	0.5	50
Low-threshold	-65	-47	-10	2.5	0.25	1.5	0.5	50

$V_{RMP}$  = resting membrane potential;  $V_{th}$  = threshold voltage;  $V_{block}$  = blockade voltage;  $\tau_{AR}$  = absolute refractory time constant;  $W_{RR}$  = relative refractory weight;  $\tau_{RR}$  = relative refractory time constant;  $W_{AHP}$  = after-hyperpolarization increment;  $\tau_{AHP}$  = after-hyperpolarization time constant.

doi:10.1371/journal.pone.0047251.t001



**Figure 1. Overview of model.** A virtual forearm with joint angle  $\theta$ , controlled by 1 flexor and 1 extensor muscle, is trained to align to a target. A *proprioceptive preprocessing* block translates muscle lengths into an arm configuration representation. Plasticity is restricted to the mapping between *sensory representation* and *motor command representation* units (dashed oval). Motor units drive the muscles to change the joint angle. The *Actor* (above) is trained by the *Critic* which evaluates error and provides a global reward or punishment signal. doi:10.1371/journal.pone.0047251.g001

$$m_{ext} = \frac{l(\theta - \theta_{min})}{\theta_{max} - \theta_{min}} \quad (1)$$

$$m_{flex} = l - m_{ext}. \quad (2)$$

Arm position updates were provided at 50 ms intervals, based on extensor and flexor EM spike counts integrated from a 40 ms window that began 90 ms prior to update time (50 ms network-to-muscle propagation delay). The angle change  $\Delta\theta$  for the joint was the difference between the flexor and extensor EM spike counts during the prior interval, with each spike causing a  $1^\circ$  rotation. P drive activity updated after an additional 25 ms delay which represented peripheral and subcortical processing. Reinforcement occurred every 50 ms with calculation of hand-to-target error. The target remained stationary during the simulation.

The RL algorithm implemented Thorndike's Law of Effect using global reward and punishment signals [17]. The network is the *Actor*. Feedforward ES  $\rightarrow$  EM AMPA weights were trained to implement the learned sensorimotor mappings. The *Critic*, a global reinforcement signal, was driven by the first derivative of error between position and target during 2 successive time points (reward for decrease; punishment for increase). As in [7], we used a spike-timing-dependent rule to trigger eligibility traces to solve the credit

assignment problem. The eligibility traces were binary-stated, turning on for a synapse when a postsynaptic spike followed a presynaptic within a time window of 100 ms; eligibility ceased after 100 ms. When reward or punishment was delivered, eligibility-tagged synapses were potentiated (long-term potentiation LTP), or depressed (long-term depression LTD), correspondingly.

Weights  $w(t)$  were updated utilizing weight scale factors,  $w_s$ :

$$w(t) = w_0 w_s(t)$$

$$w_s(t+1) = w_s(t) + \Delta w_s$$

$$\Delta w_s = \begin{cases} 1 - w_s(t)/w_s^{max} & \text{for LTP reward} \\ -w_s(t)/w_s^{max} & \text{for LTD punisher} \end{cases}$$

where  $w_s^{max}$  (5 in all simulations) is maximum weight scale factor, and  $w_0$  is the initial synaptic weight.  $w_s$  is initialized to 1.0 for all synapses and varies between 0 and  $w_s^{max}$ .

The model was implemented in NEURON 7.2 [41] for Linux and is available on ModelDB (<https://senselab.med.yale.edu/>)

**Table 2.** Area interconnection probabilities and starting weight.

Presynaptic type	Postsynaptic type	Connection probability	Synaptic weight
P	ES	0.10	8.77
ES	IS	0.43	1.90
ES	ILS	0.51	0.95
ES	EM	0.08	5.28*
IS	ES	0.44	4.50
IS	IS	0.62	4.50
IS	ILS	0.34	4.50
ILS	ES	0.35	1.25
ILS	IS	0.53	2.25
ILS	ILS	0.09	4.50
EM	IM	0.43	1.90
EM	ILM	0.51	0.95
IM	EM	0.44	4.50
IM	IM	0.62	4.50
IM	ILM	0.34	4.50
ILM	EM	0.35	1.25
ILM	IM	0.53	2.25
ILM	ILM	0.09	4.50

\*shows plastic connections, for which the initial weight is listed.  
doi:10.1371/journal.pone.0047251.t002

modeldb). One minute of simulated time took approximately 80 s of CPU time on an Intel XEON 2.27 GHz CPU.

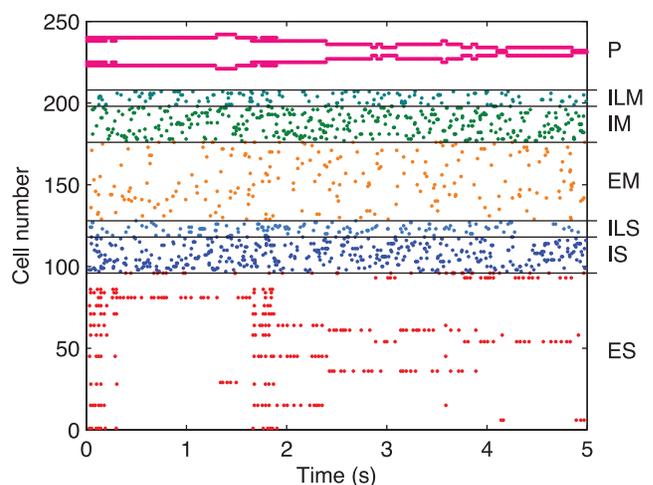
**Table 3.** Noise stimulation to synapses of the different cell types.

Cell type	Synapse type	Synaptic weight	Average rate (Hz)
IS	AMPA	4.13	300
IS	NMDA	1.50	50
IS	<i>soma</i> GABA	1.88	125
IS	<i>dend</i> GABA	1.88	125
ILS	AMPA	3.00	300
ILS	NMDA	0.38	50
ILS	<i>soma</i> GABA	1.88	125
ILS	<i>dend</i> GABA	1.88	125
EM	AMPA	3.94	300
EM	NMDA	0.75	50
EM	<i>soma</i> GABA	1.88	125
EM	<i>dend</i> GABA	1.88	125
IM	AMPA	4.13	300
IM	NMDA	1.50	50
IM	<i>soma</i> GABA	1.88	125
IM	<i>dend</i> GABA	1.88	125
ILM	AMPA	3.00	300
ILM	NMDA	0.38	50
ILM	<i>soma</i> GABA	1.88	125
ILM	<i>dend</i> GABA	1.88	125

doi:10.1371/journal.pone.0047251.t003

**Results**

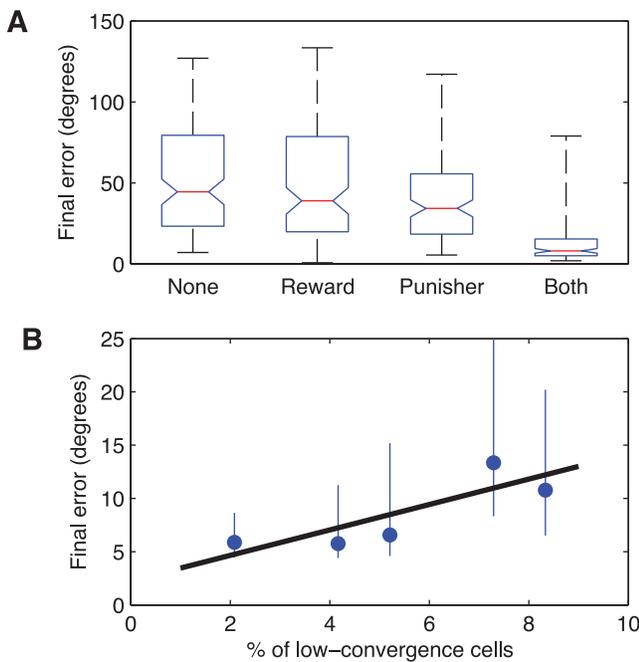
Average spiking rates (in Hz) were: P, 1.9; ES, 0.4; IS, 4.4; ILS, 2.9; EM, 0.5; IM, 4.3; and ILM, 3.1, in the absence of learning and the presence of babble noise (Fig. 2). Inhibitory cells fired faster than excitatory cells, consistent with observed rates in cortex. The top units in Fig. 2 represent the proprioceptive (P) inputs from the flexor and extensor muscles. Each of the 2 muscles stimulates 1 or 2 of the P cells to fire, with the particular cells depending on current muscle length.



**Figure 2. Baseline spiking before learning.** Raster of spikes from individual cells: ES, excitatory sensory units; IS, fast-spiking inhibitory sensory units; ILS, low-threshold inhibitory sensory units; EM, excitatory motor units; IM, fast-spiking inhibitory motor units; ILM, low-threshold inhibitory motor units; and P, proprioceptive input units.  
doi:10.1371/journal.pone.0047251.g002

We ran 500 simulations (4 learning modes: no learning, reward-only, punisher-only, reward-and-punisher×5 target angles: 0°, 35°, 75°, 105°, 135°×5 wiring random seeds×5 babbling noise input random seeds) to assess performance of the model in learning to reach for a single target. For each trial, final error was measured by the average absolute value angle difference between hand and target in the last 20 s of the 200 s trials. A Shapiro-Wilk test on all final errors for the reward-and-punisher condition found that the null hypothesis of normality could be rejected ( $p < 10^{-14}$ ), so non-parametric statistics have been used throughout. The reward-and-punisher learning algorithm showed clear superiority over reward-only or punisher-only methods (Fig. 3A;  $p < 10^{-6}$ , Kruskal-Wallis test). Final error was much less with both reward and punishment (median = 8.07, IQR = 5.10–15.23) than with the other cases (median = 38.96, IQR = 19.85–78.53 for the reward-only condition). Therefore, reward-and-punisher learning was used for all of the additional studies.

Learning with this algorithm was successful with arbitrary choice of random babble inputs. However, some randomly chosen wirings produced networks that would not learn adequately across the set of targets ( $p < 0.0003$  difference for wirings; Kruskal-Wallis test with Bonferroni correction). The difference across different network architectures was explained by noting that poorly performing networks had a relatively high proportion of poorly connected EM neurons, defined as cells with fewer than five inputs from sensory cells (Fig. 3B). These neurons would not receive adequate drive and would therefore contribute less to the dynamics. Conversely, networks with more consistent numbers of inputs per cell were more flexible and thus better at learning. In



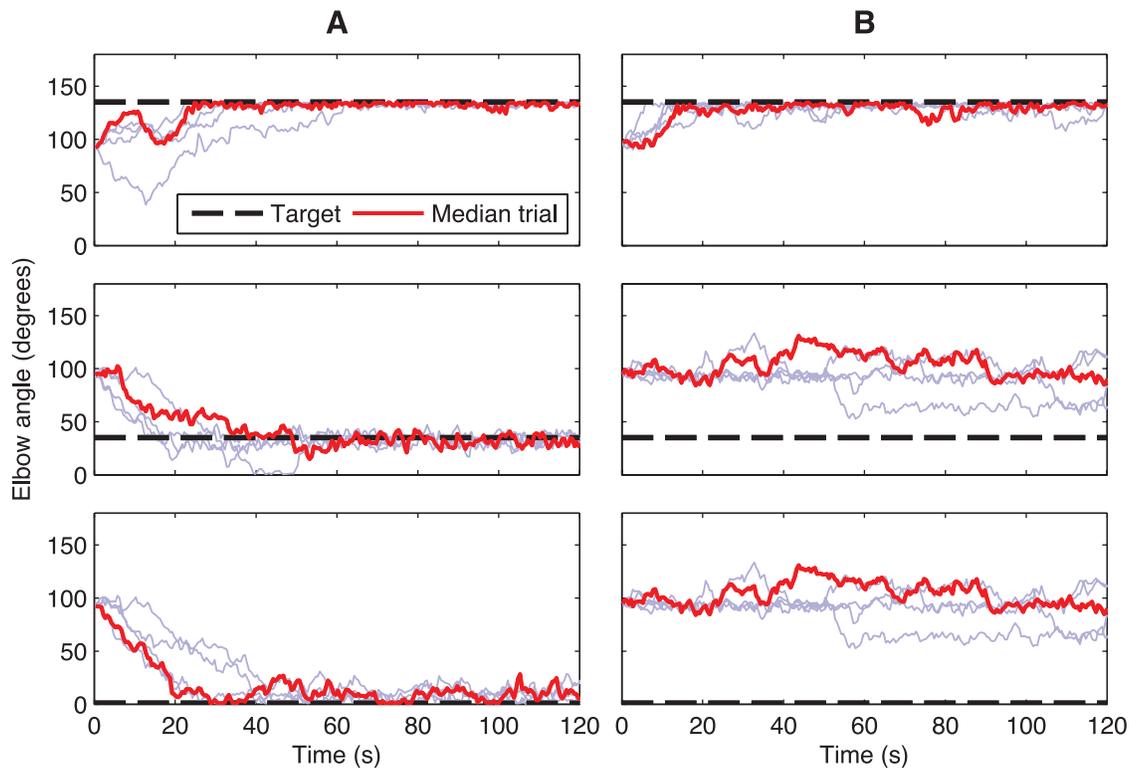
**Figure 3. Performance across learning conditions and random wiring seeds.** A. End-of-trial errors under different learning conditions: no learning, reward-only, punisher-only, reward-and-punisher;  $N = 125$  for each: 5 target angles (0°, 35°, 75°, 105°, 135°)×5 random wirings×5 random babble noise inputs. B. Learning performance as a function of the percentage of EM cells that have low-convergence, defined as having fewer than five afferent inputs. The final error appears to be strongly correlated with the proportion of low-convergence cells ( $N = 5$ ; Pearson's  $r = 0.83$ ,  $p = 0.08$ ). doi:10.1371/journal.pone.0047251.g003

addition, the wirings that produced unevenly-performing (across targets) networks often appeared to have a strong innate bias towards flexion or extension. Fig. 4 compares a wiring that produced good learning across all 5 targets (3 shown; Fig. 4A) compared to another wiring that would only learn the 2 most flexed targets (1 shown; Fig. 4B). Note that all the random babble activations in Fig. 4B produced networks that learned the flexion position (135°) and did not learn any other position, save 105° (not shown). Furthermore, these networks learned flexion more quickly than any of the networks of Fig. 4A.

The successful performance seen in Fig. 4A might have depended on metastable learning: learning due solely to ongoing reward and punishment feedback changing the proprioceptive-to-motor mapping in an *ad hoc* fashion, as opposed to a stable mapping being learned to guide the arm in the correct direction. We therefore tested for stability by assessing performance with learning turned off after training. Performance remained stable with median final error of 6.8° (IQR = 4.1–13.0;  $N = 125$  simulations), statistically unchanged from the final error with learning turned on. Stable learning suggested the development of an attractor around the target that was sufficiently deep to compensate for deviations produced by ongoing babble input.

In order to assess learning algorithm adaptability to altered environmental circumstances, we switched targets after training (targetswitch). We performed a Wilcoxon sign-rank test to compare performance on the first and second target positions and found that there was no significant difference ( $p = 0.18$ ,  $N = 50$ ). Likewise, when we performed a Wilcoxon rank-sum statistic test comparing performance on the first target sequence (Fig. 5A) vs. the second (Fig. 5B), we found no significant difference ( $p = 0.62$ ,  $N = 25$ ). This suggests that target switching performance is independent of target ordering or place in the sequence. The new target was generally learned rapidly, within 25 s after the switch. Those wirings that could not learn the separate single target trials for both first and second target could also not learn target switching between the two.

Plasticity was confined to ES→EM weights. Reward-only and punisher-only algorithms each only allowed change in 1 direction and thus produced monotonic alterations in weights. The reward-and-punisher algorithm produced a net increase in weights, but allowed weights to balance in a way not possible with only one learning direction. Separate populations of EM cells drove extension and flexion of the virtual forearm, with overall arm motion being determined by the difference between the aggregate subpopulation firing rates. To understand the net effect of learning, we examined the activity produced at each arm angles at increments of 5°. Fig. 6A shows the rotation commands that are the differences of flexor (red) and extensor (blue) EM cell aggregate firing rates for post-learning. The black vertical line indicates the target the model was trained to reach for (35°). The dotted vertical line at 25° indicates a condition where the arm angle is less than the target, and the dotted line at 65° indicates a condition where the arm angle is greater than the target. Fig. 6B shows the firing of the individual EM cells under the over-extended example (25°). Both flexion and extension EM cells learn to burst here in response to 25°, but the net firing of the flexion cells (red) wins. Fig. 6C shows the firing of the individual EM cells under the over-flexed example (65°). Here, only extension cells effectively learn to burst in response to 65°, making extension (blue) the clear winner. Learning generally caused certain cells to burst at rates up to 12 Hz in response to particular detected angles. With some notable exceptions (e.g. 10° and 70°), Fig. 6A indicates that flexion wins, as is desired, when the arm angle is less than the target, and



**Figure 4. Target learning for two model wirings.** Columns A and B show different wiring seeds, while each panel shows 5 different babble input random seeds with median final performance in red (calculated over 100–200 s; only 0–120 s shown).  
doi:10.1371/journal.pone.0047251.g004

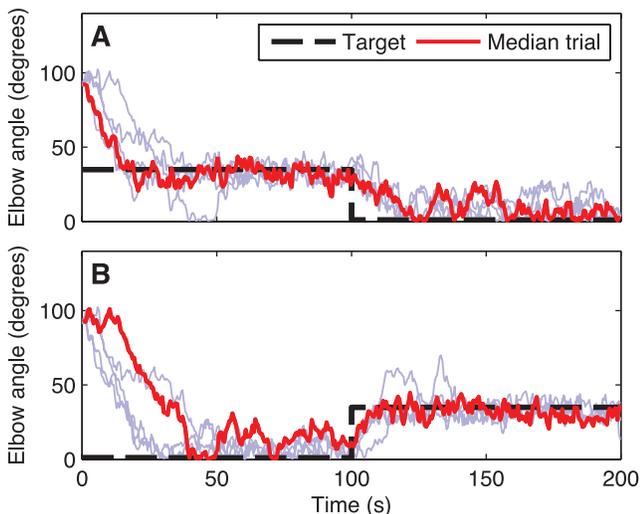
extension wins when the angle is greater. The aggregate learning effect is an attractor around the target at  $35^\circ$ .

## Discussion

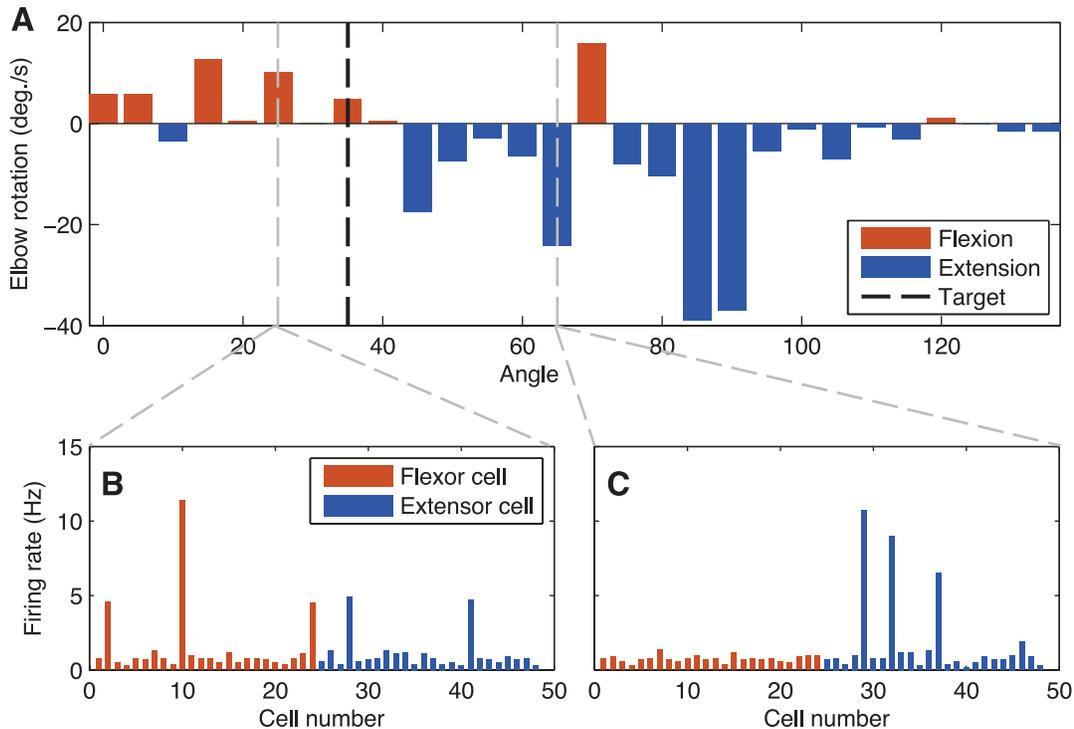
We have created a model of motor reinforcement learning for a simple target reaching task for a virtual forearm, using spiking units whose weights are trained using a reinforcement learning algorithm. Baseline spiking rates were adjusted to be comparable

with rates observed in cortical tissue (Fig. 2). Both reward and punishment together were essential for learning the task (Fig. 3A). The model was able to learn a stable attractor for the target, not merely relying on metastable, ad hoc learning for performance. The model was not only able to learn under different initial target training situations, but was able to relearn (Fig. 5). Reinforcement learning worked by shaping the collective activity of the EM cells through synaptic plasticity in the ES to EM projection (Fig. 6). This plasticity shaped EM responses to the proper mixture of ES activations to produce the desired behavior. We note that the simplicity of our one degree-of-freedom task, compared to the complexities of limb movement, reduces the scope of what can be demonstrated in this model. In particular, the network did not need to obtain stable intermediate synaptic strengths, as would be required in higher-dimensional tasks.

We predict that learning of a motor task will require both reward and punishment signals, hence both LTP and LTD in motor cortex, regulated through differential dopaminergic signaling. However, there are many additional synaptic influences in neocortex that were not included in this model. Adaptation in cell firing rate based on global synaptic input (homeostasis) or interactions between the strength of synapses (normalization) could provide alternative ways of achieving synaptic decrementation without punishment/LTD. All of frontal cortex, including M1, is innervated extensively by dopaminergic projections from the ventral tegmental area which could provide regulation of plasticity [30–32]. There is evidence that increased dopamine concentration leads to synaptic LTP and that decreased dopamine concentration leads to synaptic LTD mediated via action of D1-family receptors [42,43]. We therefore predict that dopaminergic innervation of M1 from the VTA would be required for learning, and that antagonism of D1 would be likely to impair acquiring of



**Figure 5. Serial target learning.** A.  $35^\circ \rightarrow 0^\circ$ . B.  $0^\circ \rightarrow 35^\circ$ . (blue: 4 babble traces; red: median trial; wiring from Seed 1 of Fig. 4).  
doi:10.1371/journal.pone.0047251.g005



**Figure 6. Learned EM activation for target of 35°.** A. Rotation command (differences of flexor and extensor EM rates) as a function of angle. Example firing rates for individual cells are shown at two representative arm angles: 25° (B, extended relative to target) and 65° (C, relatively flexed). doi:10.1371/journal.pone.0047251.g006

motor tasks. Various kinds of learning are impaired by disruption of dopaminergic pathways [44–48]. Parkinson disease patients, who have damage to midbrain dopaminergic nuclei, including VTA [49], have deficits reward-based learning [49–51].

Stable learned mappings can still permit rapid learning and unlearning due to shifting reinforcement conditions (Fig. 5). Similar metastable behavior is seen in the ongoing error corrections in adult bird-song production, which relies on sensorimotor integration [52]. Shifting reinforcement conditions are a typical feature of an animal's environment. Areas of the habitat once rich in food may become depleted, or once-safe areas may later be occupied by predators, making the capability for rapid unlearning and relearning of reward and punishment conditions important for survival.

Babble noise in a motor system is likely to be important in the exploration needed to drive successful animal reinforcement learning. Random motor activity provides the variation required by selection. Movements or programs can then be reinforced, consistent with a selective hypothesis [53]. This interaction of babble and learning has been most clearly demonstrated in the variability of Bengalese finch bird-song. In this species, the lateral magnocellular nucleus of the anterior nidopallium (LMAN), a part of the basal-ganglia forebrain circuit, projects to pre-motor areas which activate song production. LMAN provides a source of variability which is required for song-learning to take place via a process of random exploration and learning [54,55]. In primates, exploratory behavior has been associated with activity in anterior cingulate cortex (ACC) [56,57] and frontopolar cortex (BA 10)

[58,59], which is also connected to ACC [60]. It has been proposed that there may be a ventral striatal-to-cortical loop gating activity in one or both of these areas that mediates the onset and offset of motor babbling noise applied to the cortical actor [20].

In our present model, babble remained at the same level throughout learning. Babble thus interfered with the stability of learned mapping. This interference could be reduced by setting babble noise adaptively to reflect the current level of reward and punishment that the actor is receiving: high levels of punishment or low levels of reward should encourage babbling, whereas high levels of reward should discourage it [20]. This would reduce overall exploratory behavior, but allow it to be re-engaged during environmental change. Variability in the motor system should be maintained, yet be carefully regulated [55].

## Acknowledgments

The authors would like to thank Larry Eberle (SUNY Downstate) for Neurosim lab support; Joe Francis (SUNY) for helpful discussions; Michael Hines (Yale) and Ted Carnevale (Yale) for NEURON support; Tom Morse (Yale) for ModelDB support; and the two anonymous reviewers for their helpful comments.

## Author Contributions

Conceived and designed the experiments: GC SN CK WL. Performed the experiments: GC SN. Analyzed the data: GC. Contributed reagents/materials/analysis tools: SN GC WL. Wrote the paper: GC SN CK WL.

## References

1. Hecht-Nielsen R (1989) Theory of the backpropagation neural network. In: Neural Networks, 1989. IJCNN., International Joint Conference on. IEEE, 593–605.
2. Rumelhart D, McClelland J (1986) Parallel distributed processing: Psychological and biological models, volume 2. The MIT Press.

3. Kohonen T (1990) The self-organizing map. *Proceedings of the IEEE* 78: 1464–1480.
4. Tesauo G (1995) Temporal difference learning and TD-Gammon. *Communications of the ACM* 38: 58–68.
5. Farries M, Fairhall A (2007) Reinforcement learning with modulated spike timing-dependent synaptic plasticity. *Journal of Neurophysiology* 98: 3648–3665.
6. Florian R (2007) Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation* 19: 1468–1502.
7. Izhikevich E (2007) Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb Cortex* 17: 2443–2452.
8. Potjans W, Morrison A, Diesmann M (2009) A spiking neural network model of an actor-critic learning agent. *Neural Computation* 21: 301–339.
9. Seung H (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron* 40: 1063–1073.
10. Dan Y, Poo M (2004) Spike timing-dependent plasticity of neural circuits. *Neuron* 44: 23–30.
11. Roberts P, Bell C (2002) Spike timing dependent synaptic plasticity in biological systems. *Biol Cybern* 87: 392–403.
12. Song S, Miller K, Abbott L (2000) Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neurosci* 3: 919–926.
13. Singer W (2003) Synchronization, binding and expectancy. *The handbook of brain theory and neural networks* : 1136–1143.
14. Thorpe S, Fize D, Marlot C, et al. (1996) Speed of processing in the human visual system. *Nature* 381: 520–522.
15. VanRullen R, Guyonneau R, Thorpe S (2005) Spike times make sense. *Trends Neurosci* 28: 1–4.
16. Baker S, Kilner J, Pinches E, Lemon R (1999) The role of synchrony and oscillations in the motor output. *Experimental Brain Research* 128: 109–117.
17. Thorndike E (1911) *Animal intelligence*. New York: Macmillan.
18. Sutton R, Barto A (1998) *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press.
19. Sanchez J, Tarigoppula A, Choi J, Marsh B, Chhatbar P, et al. (2011) Control of a center-out reaching task using a reinforcement learning brain-machine interface. In: *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*. IEEE, 525–528.
20. Chadderdon G (2009) A neurocomputational model of the functional role of dopamine in stimulus-response task learning and performance. Ph.D. thesis, Indiana University. URL <http://pqdtopen.proquest.com/#viewpdf?dispub=3355003>.
21. Joel D, Niv Y, Ruppin E (2002) Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks* 15: 535–547.
22. Demiris Y, Dearden A (2005) From motor babbling to hierarchical learning by imitation: a robot developmental pathway.
23. Der R, Martius G (2006) From motor babbling to purposive actions: Emerging self-exploration in a dynamical systems approach to early robot development. *From Animals to Animats 9* : 406–421.
24. Wanjerkhede S, Bapi R (2007) Modeling the sub-cellular signaling pathways involved in reinforcement learning at the striatum. *Progress in Brain Research* 168: 193–206.
25. Magee J, Johnston D (1995) Synaptic activation of voltage-gated channels in the dendrites of hippocampal pyramidal neurons. *Science* 268: 301–304.
26. Takechi H, Eilers J, Konnerth A (1998) A new class of synaptic response involving calcium release in dendritic spines. *Nature* 396: 757.
27. Hollerman J, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci* 1: 304–309.
28. Schultz W (1998) Predictive reward signal of dopamine neurons. *Journal of Neurophysiology* 80: 1–27.
29. Ungless M, Magill P, Bolam J (2004) Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science* 303: 2040–2042.
30. Hosp J, Pektanovic A, Rioult-Pedotti M, Luft A (2011) Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning. *The Journal of Neuroscience* 31: 2481–2487.
31. Luft A, Schwarz S (2009) Dopaminergic signals in primary motor cortex. *International Journal of Developmental Neuroscience* 27: 415–421.
32. Molina-Luna K, Pektanovic A, Röhrich S, Hertler B, Schubring-Giese M, et al. (2009) Dopamine in motor cortex is necessary for skill learning and synaptic plasticity. *PLoS ONE* 4: e7082.
33. Almassy N, Edelman G, Sporns O (1998) Behavioral constraints in the development of neuronal properties: a cortical model embedded in a real-world device. *Cerebral Cortex* 8: 346–361.
34. Sporns O, Alexander W (2002) Neuromodulation and plasticity in an autonomous robot. *Neural Networks* 15: 761–774.
35. Lytton W, Stewart M (2005) A rule-based firing model for neural networks. *Int J Bioelectromagnetism* 7: 47–50.
36. Lytton W, Stewart M (2006) Rule-based firing for network simulations. *Neurocomputing* 69: 1160–1164.
37. Lytton W, Omurtag A (2007) Tonic-clonic transitions in computer simulation. *J Clin Neurophysiol* 24: 175–181.
38. Lytton W, Neymotin S, Hines M (2008b) The virtual slice setup. *J Neurosci Methods* 171: 309–315.
39. Lytton W, Omurtag A, Neymotin S, Hines M (2008) Just-in-time connectivity for large spiking networks. *Neural Comput* 20: 2745–2756.
40. Neymotin S, Lee H, Park E, Fenton A, Lytton W (2011b) Emergence of physiological oscillation frequencies in a computer model of neocortex. *Front Comput Neurosci* 5: 19.
41. Carnevale N, Hines M (2006) *The NEURON Book*. New York: Cambridge University Press.
42. Reynolds J, Wickens J (2002) Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks* 15: 507–521.
43. Shen W, Flajolet M, Greengard P, Surmeier D (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. *Science's STKE* 321: 848.
44. Faure A, Haberland U, Condé F, El Massioui N (2005) Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *The Journal of Neuroscience* 25: 2771–2780.
45. Robbins T, Giardini V, Jones G, Reading P, Sahakian B (1990) Effects of dopamine depletion from the caudate-putamen and nucleus accumbens septi on the acquisition and performance of a conditional discrimination task. *Behavioural Brain Research* 38: 243–261.
46. O'Neill M, Brown V (2007) The effect of striatal dopamine depletion and the adenosine A2A antagonist KW-6002 on reversal learning in rats. *Neurobiology of Learning and Memory* 88: 75–81.
47. Parkinson J, Dalley J, Cardinal R, Bamford A, Fehrer B, et al. (2002) Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behavioural Brain Research* 137: 149–163.
48. Smith-Roe S, Kelley A (2000) Coincident activation of NMDA and dopamine D1 receptors within the nucleus accumbens core is required for appetitive instrumental learning. *The Journal of Neuroscience* 20: 7737–7742.
49. Cools R (2006) Dopaminergic modulation of cognitive function-implications for L-DOPA treatment in parkinson's disease. *Neuroscience & Biobehavioral Reviews* 30: 1–23.
50. Frank M, O'Reilly R (2006) A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience* 120: 497.
51. Frank M, Seeberger L, O'Reilly R (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306: 1940–1943.
52. Sober S, Brainard M (2009) Adult birdsong is actively maintained by error correction. *Nature Neurosci* 12: 927–931.
53. Edelman G (1987) *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books New York.
54. Kao M, Doupe A, Brainard M (2005) Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. *Nature* 433: 638–643.
55. Tumer E, Brainard M (2007) Performance variability enables adaptive plasticity of crystallized adult birdsong. *Nature* 450: 1240–1244.
56. Shima K, Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. *Science* 282: 1335–1338.
57. Bush G, Vogt B, Holmes J, Dale A, Greve D, et al. (2002) Dorsal anterior cingulate cortex: a role in reward-based decision making. *Proceedings of the National Academy of Sciences* 99: 523.
58. Daw N, O'Doherty J, Dayan P, Seymour B, Dolan R (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441: 876–879.
59. Koehlin E, Hyafil A (2007) Anterior prefrontal function and the limits of human decision-making. *Science* 318: 594–598.
60. Mufson E, Pandya D (1984) Some observations on the course and composition of the cingulum bundle in the rhesus monkey. *The Journal of Comparative Neurology* 225: 31–43.