A learning rule that explains how rewards teach attention

Jaldert O. Rombouts[1], Sander M. Bohte[1], Julio Martinez-Trujillo[2] and Pieter R. Roelfsema[3,4,5*]

[1] Department of Life Sciences, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands.

[2]Cognitive Neurophysiology Laboratory, Department of Physiology, McGill University, Montreal, QC H3G 1Y6, Canada

[3]Department of Vision & Cognition, Netherlands Institute for Neurosciences, an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), Meibergdreef 47, 1105 BA, Amsterdam, The Netherlands.

[4]Department of Integrative Neurophysiology, Centre for Neurogenomics and Cognitive Research, VU University, De Boelelaan 1085, 1081 HV Amsterdam, Amsterdam, The Netherlands

[5]Psychiatry Department, Academic Medical Center, Amsterdam, The Netherlands

*Correspondence to: p.roelfsema@nin.knaw.nl

Keywords: attention; reinforcement learning; neural networks; neuronal plasticity, top-down

Abstract

Many theories propose that top-down attentional signals control processing in sensory cortices by modulating neural activity. But who controls the controller? Here we investigate how a biologically plausible neural reinforcement learning scheme can create higher order representations and top-down attentional signals. The learning scheme trains neural networks using two factors that gate Hebbian plasticity: (1) an attentional feedback signal from the response-selection stage to earlier processing levels and (2) a globally available neuromodulator that encodes the reward prediction error. We demonstrate how the neural network learns to direct attention to one of two coloured stimuli that are arranged in a rank-order (Lennert & Martinez-Trujillo, 2011). Like monkeys trained on this task, the network develops units that are tuned to the rank-order of the colours and it generalizes this newly learned rule to previously unseen colour combinations. These results provide new insight into how individuals can learn to control attention as a function of reward contingency.

Introduction

Our perception is highly selective. We mainly register information that pertains to our goals while ignoring the rest. Consider, for example, a tennis player waiting for the return of the ball. He focuses on the posture and motion of the opponent and on how to hold the racquet, but he neither perceives other people on the court nor the surrounding advertisements. Through training he has learned to focus attention on the visual information that matters for the next hit. How did he learn to specifically attend to the visual features that matter?

In many circumstances learning depends on rewards or punishments. Winning and losing points during practice games is an incentive for learning how to play tennis. Reinforcement learning (RL) theories provide a useful framework for understanding how feedback from the environment in the form of rewards and punishments shapes behavioural performance (Sutton & Barto, 1998). Researchers have made substantial progress in RL theories, and there are influential theories about how the brain implements RL (Bromberg-Martin, Matsumoto, & Hikosaka, 2010; Dayan & Balleine, 2002). Furthermore, other influential theories have addressed how attention influences neuronal activity in visual cortical areas (Bundesen, Habekost, & Kyllingsbæk, 2005; Desimone & Duncan, 1995; Reynolds & Chelazzi, 2004; Roelfsema, 2006). However, with a few exceptions (e.g. Whitehead & Ballard, 1991), previous theories have not yet addressed the question of how the brain learns to direct attention to those features that matter. In the present study we will investigate a new, biologically plausible RL-scheme with the aim to train a neural network to control attention. Our approach is inspired by recent findings about the influence of rewards on attention in experimental psychology and also by neurophysiological findings on how attention influences the representation of visual information in the brain.

In recent years, researchers in experimental psychology obtained many new insights into how rewards influence the deployment of attention, as documented in this special issue of *Visual Cognition*. Visual stimuli that are associated to high reward are likely to attract more attention at a later point in time than stimuli associated with lower reward (Chelazzi, Perlato, Santandrea, & Libera, 2013) (B. A. Anderson, Laurent, & Yantis, 2011; Hickey, Chelazzi, & Theeuwes, 2010; Libera & Chelazzi, 2009; Raymond & O'Brien, 2009). In many tasks the reward-contingencies determine what information is relevant and what information is not, which makes it evident that reinforcers should influence attention.

Neurophysiological studies have demonstrated that areas of the parietal and frontal cortex selectively represent task-relevant information, and the pairing of stimuli with rewards and punishments should therefore favour the representation of these stimuli (Duncan, 2010; Gottlieb & Balan, 2010). Most theories of attention state that these neurons in frontal and parietal cortex provide a top-down signal that influences the representation of stimuli in the visual cortex (Corbetta & Shulman, 2002; Desimone & Duncan, 1995; Miller & Cohen, 2001). Indeed, the neuronal responses elicited by attended objects are enhanced in many, if not all, areas of visual cortex (Reynolds & Chelazzi, 2004; Treue & Maunsell, 1996), and attentional selection processes can be monitored even at the level of the primary visual cortex (area V1) (Roelfsema, 2006). Because the required top-down attentional control signals depend on the precise task demands, they should be strongly influenced by RL. However, the mechanisms that allow reward signals to shape these attentional top-down control signals are not well understood.

Interestingly, neuronal activity evoked by stimuli associated with high rewards is also stronger in visual and association cortex than activity evoked by stimuli

associated with less reward (Louie, Grattan, & Glimcher, 2011; Pastor-Bernier & Cisek, 2011; Serences, 2008; Stănişor, van der Togt, Pennartz, & Roelfsema, 2013). Although some have proposed that the effects of attentional selection and reward expectancy differ at the neuronal level (Louie et al., 2011; Platt & Glimcher, 1999), a recent study demonstrated that V1 neurons modulated by reward expectancy are also modulated by attention and with a similar timing (Stănişor et al., 2013). The latter suggests that both response modulations are caused by a common top-down signal reaching the visual cortex driven by both reward expectation and selective attention (Maunsell, 2004).

To gain more insight into how reinforcers can influence the deployment of attention, we will here train a neural network model to carry out a non-linear attentional control task that has been studied in monkeys by Lennert & Martinez-Trujillo (2011). We will train the network with a new learning rule called AuGMEnT (Rombouts, Bohte, & Roelfsema, 2012; in press), which incorporates two factors that are known to modulate synaptic plasticity (Roelfsema & van Ooyen, 2005; Roelfsema, van Ooyen, & Watanabe, 2010). The first factor is a reward-prediction error that codes whether the outcome of an action is better or worse than expected. It is thought that such a reward prediction error is broadcasted throughout the brain by the release of neuromodulators, such as dopamine or serotonin, so that it is available at many synapses and can influence their plasticity (Liu et al., 2014; Schultz, 2002). The second signal is an attentional top-down signal from the response selection stage to earlier processing levels. The learning rule enforces that this top-down signal highlights the subset of synapses that are responsible for the action that was chosen by the network and this signal thus determines which synapses are sensitive to the neuromodulators that determine the changes in synaptic strength. Overall, the learning rule incorporates four

signals that are all available locally, at the synapse: (i) presynaptic activity; (ii) postsynaptic activity; (iii) the globally released neuromodulatory signal and (iv) activity of feedback connections from the response selection stage.

In the task of Lennert and Martinez-Trujillo (2011) the monkeys first had to direct their gaze to a fixation mark in the centre of a display flanked by two moving random dot patterns (RDPs) made up of grey dots (Figure 1A). After a delay, both RDPs changed colour. These colour changes were an attentional cue for the monkeys, indicating which of the patterns was target and which one distracter (e.g., green was the target and red the distracter). The monkeys' task was to respond to a brief change in the motion direction of the target pattern by releasing a button while ignoring a similar change in the distracter's direction. The crucial design feature of the task is that that were six possible colours, which were organized according to their rank in an ordinal scale (red < orange < yellow < green < blue < purple; the original study used different colours which did not map onto the spectrum in this orderly manner). The monkeys had to attend to the pattern with the highest colour rank (the target) and ignore the pattern with the lowest rank (the distracter). They only received a juice reward for responses to the target direction change, whereas trials were aborted without reward if they responded to the distracter change. Note that this task is non-linear, because the colour cues determine whether a response is required to the left motion cue so that the right motion cue should be ignored, or whether the contingency is reversed. The monkeys learned the order relationships by trial and error, which is remarkable because they saw only two coloured patterns at the same time. They generalized the rule to infer the relative rank of new colour pairs that they had not seen during training (i.e., transitive knowledge).

Once the monkeys had learned the task, Lennert & Martinez-Trujillo (2011) recorded the activity of neurons in the dorsolateral prefrontal cortex. They observed that the firing rate of a substantial fraction of the cells coded the location of the target pattern. Some of these neurons increased their firing rate if the target pattern was on the left side of the display, whereas others increased their response if the target pattern was on the right. Moreover, the strength of the attentional control signal depended on the distance between the ranks of the target and distracter colours. The signal was strongest if the distance between the ranks was high and weaker for colours with adjacent ranks (distance effect).

Although the study of reinforcement learning was not the aim of Lennert & Martinez-Trujillo (2011), as rewards were only used as incentive for the monkeys to do the task, this study does allow us to address the central theoretical issues that we wish to investigate: how do brain structures control attention and how do they optimize this control while animals learn a task by reinforcement learning? More specifically: (1) which mechanism can cause neurons in prefrontal cortex to encode the rank order of the colours? (2) how do these control signals ensure that the monkey only responds to changes in the direction of the target? and (3) how do the monkeys generalize the rule to new colours that they have not seen during training?

To address these questions, we exploited a new versatile learning scheme called AuGMEnT (Attention-Gated MEmory Tagging) that can train neural networks to perform many of the tasks that have been used in monkey studies, including tasks that require decision-making, non-linear sensory-motor mappings, working memory and categorization (Rombouts et al., 2012; in press). The network aims to learn the value of the successive actions that need to be taken during a trial, such as holding or releasing a button. These action values correspond to the expectancy of obtaining a

reward at the end of the trial, and the representation of these action values allows the model to make the optimal choice at every time step during a trial.

A unique feature of the learning rule is that neuronal plasticity makes feedforward and feedback connections reciprocal, in accordance with anatomical and neurophysiological findings (Felleman & Van Essen, 1991; Mao et al., 2011). When the model learns to select actions based on the relevant features, the units coding these features start receiving attentional feedback from the response selection stage. We show that a model trained with AuGMEnT can indeed learn the attentional control task, and that the behaviour of the model is similar to that of monkeys. We further show that the model explains the formation of rank-difference tuning and how trial-and-error learning can shape attentional top-down signals.

Methods

Model

We have described the AuGMEnT learning rule in previous work (Rombouts et al., 2012; in press), but we have not used it so far to study how trial-and-error learning shapes top-down attentional selection signals. In these previous studies we outlined the theory behind the model and how it optimizes network performance. We also demonstrated how AuGMEnT allows networks to learn non-linear sensory-motor transformations and decision-making tasks. We will here summarize the key features of the model, which are necessary to understand learning of the attentional control task.

An important property of AuGMEnT is that it can train a two-layer neural network to perform many tasks, by simply varying the input stimuli and the reward contingency. In

the present study, we used the same network topology as in our previous work (Figure 1B) to understand how rewards influence attentional control. The model is a two layer neural network that learns to predict action values (also known as Q-values, (Sutton & Barto, 1998)) for the different actions that it can take (Figure 1B). Thus, when a stimulus is presented to the input layer, the model's task is to propagate activity from the input layer to the association layer and then to the Q-value layer to compute the value of the different actions that the model can take. Phrased more formally, there is a Q-value unit for every possible action *a* and this unit aims to represent the (discounted) expected reward for the remainder of a trial if the network selects an action *a* in the current state *s*:

$$Q^{\pi}(s,a) = E_{\pi}[R_t \,|s_t = s, a_t = a\,], \text{ with } R_t = \sum_{p=0}^{\infty} \gamma^p r_{t+p+1}\,, \quad (1)$$

where $E_{\pi}[.]$ is the expected value of the sum of discounted future rewards R_t, given current action-selection policy π and where $\gamma \in [0,1]$ determines the discounting of future rewards *r*. Discounting means that rewards in the distant future are considered less valuable than rewards that can be earned at earlier time points.

Learning is guided by a neuromodulatory signal that represents the SARSA temporal difference error δ:

$$\delta(t) = r(t) + \gamma Q'(t) - Q(t-1), \quad (2)$$

where $r(t)$ is the scalar reward observed on time step t. SARSA is a method from the RL literature (Rummery & Niranjan, 1994; Sutton & Barto, 1998) that considers transitions from one state-action combination to the next while evaluating the reward (hence SARSA; State-Action Reward State-Action). This δ is positive if the outcome of an action was better than expected and negative if it was worse. For example, if the Q-value of an action taken on time-step *t* equals 0.8, the network expects 0.8 units reward for the remainder of the trial. If the network at time *t*+1 selects the next action with a

value of 0.9, the action at time *t* turned out to be better than expected (δ=0.1) and the network updates the synapses to increase the value of the action that was taken at time *t*.

The sensory layer ('stimuli' in Figure 1B) of the network represents the current stimulus with three different unit types: Instantaneous ($x_i(t)$) units, and On (+, $x_i^+(t)$) and Off (-, $x_i^-(t)$) units, so that each sensory input $s_i(t)$ is encoded by three different types of input units:

$$\begin{aligned} x_i(t) &= s_i(t), \\ x_i^+(t) &= [s_i(t) - s_i(t-1)]_+, \\ x_i^-(t) &= [s_i(t-1) - s_i(t)]_+, \end{aligned} \tag{3}$$

where $[\cdot]_+$ is a threshold operator that leaves positive inputs unchanged but returns 0 for negative inputs. Thus, instantaneous units code for the current sensory input, whereas On-units are active for only one time-step if a feature has just appeared and Off-units when it disappeared.

The association (middle) layer of the network (middle in Figure 1B) is equipped with two different types of units: regular units and memory units. The activity of regular units depends on the current activity of units in the input layer, whereas memory units exhibit persistent activity so that they can represent working memories of stimuli presented earlier, as are found in for instance in prefrontal and parietal cortex (Funahashi, Bruce, & Goldman-Rakic, 1989; Gnadt & Andersen, 1988). We will first describe the activity of regular units before we describe the activity of the memory units.

Instantaneous units x_i in the input layer project to the regular association units via synaptic weights v_{ij}^R (with v_{0j}^R is a bias weight) (Figure 1B). Their activity y_j^R is determined by:

$$inp_j^R(t) = \sum_i v_{ij}^R x_i(t)\ ,$$
$$y_j^R(t) = \sigma\left(inp_j^R(t)\right) \equiv 1/(1+\exp\left(\theta - inp_j^R(t)\right)\ \ , \tag{4}$$

where $\sigma(.)$ is a non-linear activation function (squashing function) and we note that our results generalize to other forms of this activation function. The activation function of the memory units is similar. The on (+) and off (-) units in the sensory layer project to the memory units via synaptic weights v_{im}^+ and v_{im}^- and their activation is determined as:

$$inp_m^M(t) = inp_m^M(t-1) + \sum_i \left(v_{im}^+ x_i^+(t) + v_{im}^- x_i^-(t)\right),$$
$$y_m^M(t) = \sigma\left(inp_m^M(t)\right)\ , \tag{5}$$

so that their activity also depends on features that have appeared or disappeared during earlier time steps in the trial. The instantaneous units in the input layer do not project to the memory units to prevent the integration of a constant input, which would give rise to ramping activity of the memory units.

Finally, the regular and memory association units project to the output/motor layer via synaptic weights w_{jk}^R (with w_{0k}^R as a bias weight) and w_{mk}^M, respectively, to give rise to a set of Q-values q_k for the different actions k that the model can take so that:

$$q_k(t) = \sum_m w_{mk}^M y_m^M(t) + \sum_j w_{jk}^R y_j^R(t). \tag{6}$$

Thus, when activity has been propagated to the output layer, this layer encodes a set of Q-values, one for every action. Then a stochastic winner take all (WTA) competition determines the action that the network will perform. With high probability $(1-\epsilon)$ the greedy action (with highest Q) is selected, but with a small probability ϵ the winning

action is determined by sampling from the Boltzmann distribution to allow the exploration of other actions:

$$P_B(k) = \frac{\exp(q_k)}{\sum_{k\prime} \exp(q_{k\prime})}, \tag{7}$$

where $P_B(k)$ is the probability that action k is selected. After selecting an action a, the activation in the Q-value layer becomes $z_k = I_{ka}$, where I_{ka} is an identity function returning 1 if $k = a$ and 0 otherwise. In other words, only the winning output unit *a* has activity 1 and the activity of the other units in the output layer becomes zero. The output layer then informs the rest of the network about the selected action via feedback weights w'^{M}_{km} to memory units and w'^{R}_{kj} to regular association units (dashed lines in Fig. 1B), and the interaction of the feedback and feedforward activations is used to constrain synaptic plasticity to those synapses that were actually involved in selecting the action. The network creates synaptic traces and synaptic tags that determine synaptic plasticity. The traces signal that a synapse has been active, whereas tags signal that the synapse was involved in the selection of an action, so that the synapse is going to be held responsible for the outcome of this action. Synaptic traces on synapses from regular and memory units to the output layer are proportional to the strength of the afferent signal through these synapses:

$$\begin{aligned} \mathrm{sTrace}^{R}_{jk}(t) &= y_j(t), \\ \mathrm{sTrace}^{M}_{mk}(t) &= y_m(t). \end{aligned} \tag{8}$$

These traces are a prerequisite for plasticity because tags can only form on those synapses that contain a trace. Tags in the output layer form only on synapses onto the winning output unit that was selected for an action, and they then decay exponentially:

$$\begin{aligned} \mathrm{Tag}^{R}_{jk}(t+1) &= \lambda\gamma \mathrm{Tag}^{R}_{jk}(t) + sTrace^{R}_{jk}(t) z_k(t), \\ \mathrm{Tag}^{M}_{mk}(t+1) &= \lambda\gamma \mathrm{Tag}^{M}_{mk}(t) + sTrace^{M}_{mk}(t) z_k(t), \end{aligned} \tag{9}$$

The parameter $\lambda \in [0,1]$ determines the rate of decay of tags (in RL these tags are called eligibility traces (Sutton & Barto, 1998)). The strength of the tag determines the degree of plasticity of the synapse.

The updates for the synapses v_{ij}^R between the input and association layer have a similar form:

$$\begin{aligned} \mathrm{sTrace}_{ij}^R(t) &= x_i(t), \\ \mathrm{Tag}_{ij}^R(t+1) &= \lambda\gamma \mathrm{Tag}_{ij}^R(t) + \mathrm{sTrace}_{ij}^R(t) fb_j^R(t), \\ &= \lambda\gamma \mathrm{Tag}_{ij}^R(t) + \mathrm{sTrace}_{ij}^R(t)\, \sigma'\big(inp_j^R(t)\big) w_{aj}'^R, \end{aligned} \tag{10}$$

Note that the formation of tags here depends on $fb_j^R(t)$, which is a shorthand for the attentional feedback signal that originates from the winning output unit a and arrives at unit j through the feedback connections $w_{aj}'^R$, and σ' is the derivative of the activation function with respect to its input, which has the convenient form $\sigma'\big(inp_j^R\big) = \sigma\big(inp_j^R\big)\Big(1 - \sigma\big(inp_j^R\big)\Big)$. Thus, only synapses v_{ij}^R that receive feedback from the response selection stage are rendered plastic because they form tags. The updates for the synapses $v_{im}^{+/-}$ onto memory units are similar:

$$\begin{aligned} \mathrm{sTrace}_{im}^{+/-}(t) &= \mathrm{sTrace}_{im}^{+/-}(t-1) + x_i^{+/-}(t), \\ \mathrm{Tag}_{im}^{+/-}(t+1) &= \lambda\gamma \mathrm{Tag}_{im}^{+/-}(t) + \mathrm{sTrace}_{im}^{+/-}(t) fb_m^M(t), \\ &= \lambda\gamma \mathrm{Tag}_{im}^{+/-}(t) + \mathrm{sTrace}_{im}^{+/-}(t)\, \sigma'(inp_m^M(t)) w_{am}'^M, \end{aligned} \tag{11}$$

where the critical difference is that synaptic traces to memory units accumulate, i.e. they reflect the total input provided by a synapse during the trial, whereas all other traces disappear after a single time step. After executing action a with expected value q_a and updating the traces and tags as specified above, the network makes a transition to a new state in the environment and it selects a new action a' with associated Q-value $q_{a'}$ on the next time step. The network may also receive a reward r during this transition. After

the transition, a neuromodulatory substance (e.g. dopamine) is globally released in the network, which encodes the SARSA prediction error $\delta(t)$. The concentration of the neuromodulator depends on the difference between successive Q-values, taking the discounting as well as the reward *r* into account:

$$\delta(t) = r(t) + \gamma\, q_{a'}(t) - q_a(t-1). \tag{12}$$

This prediction error is positive if the outcome of the previous action is better than expected and negative if it is worse, and it is also positive if the network experiences a transition to a higher Q-value but does not receive an immediate reward. Synaptic plasticity in the network is then simply determined by the interaction of this neuromodulatory substance with the tagged synapses as:

$$\begin{aligned} \Delta w_{jk}^{R} &= \beta\delta(t) Tag_{jk}^{R}(t);\ \Delta w_{lk}^{M} = \beta\delta(t) Tag_{lk}^{M}(t), \\ \Delta v_{ij}^{R} &= \beta\delta(t) Tag_{ij}^{R}(t);\ \Delta v_{im}^{+/-} = \beta\delta(t) Tag_{im}^{+/-}(t), \end{aligned} \tag{13}$$

where β determines the learning rate. Feedback weights are updated in the same manner. As was mentioned in the introduction, the learning rule is neurobiologically plausible because the factors that determine plasticity are the pre- and postsynaptic activity, the "attentional" feedback signal from the response selection stage and the neuromodulator coding for $\delta(t)$, signals that are all available locally, at the synapse.

When the network reaches the end of a trial the γ parameter used in the computation of the reward prediction error is set to 0 for the corresponding time-step and the dynamic parameters (i.e. tags, synaptic traces, unit activations) in the network are cleared. It can be shown that the above learning rules change the synapses in the network as to minimize the SARSA temporal difference prediction errors by stochastic gradient descent, and that AuGMEnT can be seen as a biologically plausible implementation of the SARSA(λ) learning algorithm, extended with a working memory (Rombouts et al., 2012). The non-linear units of the association layer allow the

network to learn non-linear mappings from sensory stimuli onto Q-values, as is essential in the present attentional control task. Figure 1C provides a graphical summary of the learning mechanism described above.

In order to investigate how AuGMEnT can train a network to control attention, we constructed the simplest possible network that captures the essentials of the attentional control task based on colour ranking. We equipped the network with a retinotopically organized sensory layer (Figure 1B, top) with binary neurons, i.e. neurons that were either active or silent, representing the seven possible colours on the left side and on the right side. We additionally included a binary neuron that represented the change in motion direction on each side, which the model had to report (for targets) or to ignore (for distracters) and a binary unit for the fixation mark in the middle. As in our previous work (Rombouts et al., 2012; in press), the association layer was equipped with three regular units and four memory units. We previously demonstrated that AuGMEnT can also be used to train networks with many more units in the association layer. The action layer of the network had neurons that coded for two different actions: one to hold a response button and the other to release the button. For all results reported here we used the following parameters: $\beta = 0.35, \lambda = 0.40, \gamma = 0.9$ and $\epsilon = 0.025$ and $\theta = 2.5$. We note that the results below do not critically depend on the precise value of these parameters – we found AuGMEnT to be robust across a wide range of parameters and in a large variety of tasks (Rombouts et al., 2012).

Model of the attentional control task

The task (Figure 1A) was modelled as a sequence of discrete time steps. Every trial started with an empty screen, shown for one time step. Then we presented the fixation mark flanked by the grey patterns. The model had to "press" the response button within

ten time steps, otherwise the trial was terminated without reward. If the model "held" the button for two time-steps, the colours of the stimuli changed, with the target changing to a colour with a higher rank than the distractor. If the target stimulus changed direction (i.e. when the binary 'change' neuron on the corresponding side became active) the model had to "choose" the release action within eight time steps to receive a "reward" of 1.5 units. However, if the distracter stimulus changed direction, the model had to hold the button for two additional time steps, after which the target stimulus would briefly change, indicating that the model could release the button to obtain the reward. As in our earlier work (Rombouts et al., 2012), we used a shaping strategy to encourage the model to learn to hold the button by giving a small reward (hold-reward) of 0.2 units when the model held the button for two time-steps after the fixation mark turned on. We note that this shaping strategy is not essential for the learning of the task, but that it speeds up the learning process, as in animal learning (Krueger & Dayan, 2009).

Model training

We carried out two separate sets of simulations. In the first set of simulations we investigated the generalization performance of the network by training it on a subset of all colour combinations and then testing performance for colour combinations that had not been presented during training. This allowed us to investigate whether AuGMEnT can generalize to unseen combinations of colours, as monkeys did. In the second set of simulations we presented all colour combinations from the start of training. The monkeys in (Lennert & Martinez-Trujillo, 2011) were trained for 3-5 months on this full version of the task, and these simulations allowed us to investigate the neuronal

tuning that develops after significant experience in the colour task.

Test of generalization

The first training scheme was developed to study if the model could generalize the colour-ranking scheme to unseen combinations of colours. We used the same shaping scheme as used to test the generalization performance of the monkeys in the (Lennert & Martinez-Trujillo, 2011) study. The models were first trained on the colour-pairs 'green-blue', 'yellow-green' and 'yellow-blue' (rule: yellow < green < blue). The ordering of the colours and the stimulus (target or distracter) where the first direction change occurred were randomized. We will use the term 'respond trial' for a trial where the target stimulus changed first, and 'ignore trial' for a trial where the distracter stimulus was the first to change. After the model learned the task (see below) we sequentially trained the model on unseen colours, first training on the combination 'red-yellow'. In combination with the initial set, the relative ranking of the red colour could be inferred from the 'red-yellow' pair, because red < yellow, and yellow < green < blue. If the model inferred this ordering, it should learn the other combinations with the red colour ('red-green' and 'red-blue') more easily. We repeated this training scheme for the 'orange' colour, first training on 'orange-yellow', and then testing learning speed for 'orange-green', 'orange-blue' and 'red-orange' simultaneously. We considered that learning was complete when the model made 85% correct choices in the last 100 trials of each colour pair. All networks learned the task within a median of 1,800 trials.

Training on all colour-pairs

We also evaluated the model's performance when it was immediately exposed to all colour pairs, without shaping except for the small hold-reward. The locations of colours and the distracter/target trials were generated uniformly at random. We considered

learning successful if the models made more than 85% correct choices in the last 100 presentations of all possible colour pairs, collapsing over the two possible locations of the two colours and across respond and ignore trials. All models learned this task in less than 5,200 trials.

Results

We investigated the learning behaviour of AuGMEnT using two different training schemes as explained in the methods section above. In the original study, the monkeys were either trained on a version of the task that tested their generalization to new colour combinations or in the full task with all colour combinations. In accordance with these two training schemes, we used a first scheme to test if a neural network trained with AuGMEnT would generalize the rule to new colour-pairs, and a second scheme to study the behaviour of networks that have been trained with all colour combinations.

Generalization to new colour combinations

In the first version of the task we trained 100 networks (i.e. repetitions with different initializations of the network weights) to test generalization (Methods), aiming to record the decisions that the models made during each stage of the training procedure. In the first phase, we trained the networks to discover the general rule with three basic colour pairs, ‘green-blue’, ‘yellow-green’ and ‘yellow-blue’. The only feedback that the network received about its performance was the small hold-reward if it held the response button for two time-steps and the large reward if it responded to the motion change on the relevant side. In spite of this limited feedback about their performance, all networks learned these patterns within a median of 1,800 trials, which is fast when

compared with learning by the monkeys.

Figure 2 shows the distributions of the number of mistakes that the models made in each learning stage for the different colour pairs, including trials where the model released the button prematurely, before the first change in the direction of the moving patterns. After the models had learned the ranking of green, blue and yellow they rapidly generalized to previously unseen colour pairs. Specifically, 'red-yellow' was clearly the most difficult to learn, because the colour red was introduced for the first time and its rank relative to yellow was unknown. Once the models had learned that red had a lower rank than yellow, the subsequent transfer to 'red-green' and 'red-blue' was easy, consistent with the hypothesis that the model could exploit the order relationship, as yellow was lower in rank than green and blue, so that red<yellow implies red<green and red<blue. We subsequently introduced the orange colour, pairing it with yellow. The model quickly learned that orange had a lower rank than yellow and it subsequently made only few errors with 'orange-green' and 'orange-blue'. The error rate increased slightly for the last colour pair ('orange-red'), but this also occurred when a monkey was trained on this task. This phenomenon can be explained because the relative rank of orange and red is undetermined when the model has learned that orange<yellow and red<yellow. The general learning pattern of the networks follows the monkeys' behaviour (white circles in Figure 2). Indeed, the Spearman rank correlation between the median network performance and the monkey's performance was 0.86 ($P < 0.012$). Thus, these simulations indicate that neural networks trained with AuGMEnT generalize a colour ranking scheme to unseen combinations of colours, and that the pattern of errors is similar to that shown by a monkey when trained on the same task.

Full task

We next investigated the behaviour of AuGMEnT on the full task with all possible combinations of colour stimuli, which was also used to train the monkeys. For these simulations we trained an additional 100 networks. They all managed to learn the task to criterion within a median of 2,800 trials, which is fast compared to the monkeys who took about 3-5 months of training. Note that the learning of the full task was slower than in the generalization task (1,800 trials). This slightly slower learning process can be explained by the fact that the purple (highest rank) colour was not included in the generalization task. Furthermore, the generalization task included many examples with adjacent ranks, which is helpful if the task is to infer a rank order (Krueger & Dayan, 2009).

Model accuracy as function of colour distance

When a task requires the comparison of stimuli that are "close" in rank, humans and animals tend to require a longer time to reach a decision and make more errors (Dehaene, Dehaene-Lambertz, & Cohen, 1998). Also the monkeys that were trained on the colour-rank task exhibited a clear effect of the difference in rank between the colour stimuli on the error rate (Figure 3A) (Lennert & Martinez-Trujillo, 2011).

To investigate whether the networks trained with AuGMEnT exhibit a similar sensitivity, we recorded all errors made by the networks during training, for colours separated by distance of 1, 2, or 3 on the colour scale. Figure 3B shows that networks trained with AuGMEnT exhibited a similar distance effect (one-way analysis of variance-ANOVA, Kruskal-Wallis post-hoc test, $H = 45.76, P < 0.001$). Thus, the neural networks captured many aspects of the behavioural performance of the monkeys. To examine how the networks learned to focus their attention on the side of the colour

with highest rank, we next examined the activity of the units in networks trained to perform the full task.

Activity of the units in a trained network

To obtain a first intuition of how the trained networks solve this task, Figure 4 shows the activity of two memory units in the association layer and also the activity of the *Q*-value units in the output layer of an example network, for all trial types with a green and orange stimulus. The unit on the top (light blue trace) had a stronger response when the target colour was on the left (compare the first and third column) whereas the unit in the middle row (grey trace) had a stronger response when the target colour was on the right. It can be seen that motion stimuli also had strong effects on the units' activity level; for instance the blue "left" unit received excitatory input from a motion change on the left (M1 in the first column of Figure 4 and M2 in the second column). Similarly, the grey "right" unit was excited by the right motion stimulus. Examination of the activity of Q-value units (lower row in Fig. 4) revealed that the Q-value of the "Hold" action is higher than the activity of the "Release" action, until the moment where the motion stimulus is presented on the side that needs to be monitored by the model. This activity pattern of the Q-value units follows from the fact that this example network had learned to hold the lever until the motion change occurred on the relevant side. When we examined the activity of many networks, we found that memory units had a very strong tuning to the difference in rank between the two colours. Specifically, their activity increased or decreased monotonically with the difference in rank between the two colours. In the next section, we provide a detailed analysis of rank-difference tuning in the trained networks.

Tuning to the difference in rank between colours

We found that the sensitivity of the model to the difference in the rank of the colours was mainly expressed in the synaptic weights from the On- and Off-cells in the input layer onto the memory units in the association layer. To assess the rank-difference tuning of these memory units in more detail we analysed the synaptic weights from the input layer units that coded the colours on the left and right. Figure 5A shows the input connections of the two memory units of the network of Figure 4. There was a clear relation between synaptic weights and the rank and location of the two coloured stimuli. The unit on the left of Figure 5A (light blue traces in Figure 4) had positive weights to colours of increasing rank on the left and negative weights for colours of increasing rank on the right. The unit on the right (grey traces in Figure 4) had the opposite tuning and prefers stimuli with a higher rank on the right. The panels in Figure 5B show the amount of input to these two memory units for all colour combinations, as determined by the weights in Figure 5A. The unit on the left of the figure received strong input if the higher ranked stimulus was on the left, and its activity depended on the distance in rank between the two stimuli as soon as the colour cues were presented. The unit on the right of the figure had the opposite tuning to colour combinations.

In order to quantify the prevalence of this gradual opposite left-right tuning to the colour rank across units in the association layer of all trained networks, we computed linear regression coefficients ($a_{l/r}$) between colour ranks ($R_{l/r}$) and synaptic weights w from the left and right (l/r) retinotopic location:

$$w(R_l) = a_l R_l + b_l,$$

$$w(R_r) = a_r R_r + b_r,$$

Figure 5C shows the regression coefficients for left and right stimuli ($a_{l/r}$). It can be seen that many memory units exhibited a tuning similar to that of the two units

illustrated in Figure 5A. Units with positive weights from stimuli with a high rank on one side almost invariably also had positive weights from stimuli with a low rank on the other side.

The memory units of the association layer fell into three groups: “left” (green in Figure 5C), “right” (red), and “non-tuned” (black) units. We labelled the units by performing k-Means clustering (assuming three clusters), and used this labelling to investigate how the networks solved the task. This analysis revealed that most networks had one unit preferring “left” stimuli (1.12 ± 0.35; mean number of units $\pm$ s.e.m) and another one preferring “right” stimuli (1.10 ± 0.33), while the remaining units fell into the grey category of Figure 5C without clear tuning (1.78 ± 0.48).

In monkeys, task-relevant neurons encode information about the rank-difference of stimuli and thereby which of the two stimuli should be monitored for a motion change. One example neuron is shown in Figure 6A. This neuron increased its response if the target stimulus appeared on the preferred side of the neuron and in particular if the difference in colour rank between the preferred and non-preferred was large. A similar effect was observed at the population level when cells were aligned according to their preferred side, which differed across neurons (Figure 6B, bottom). When we examined the activity of the top (light blue) ‘left’ coding unit in response to colour pairs with different rank-distances we found that the unit’s activation after the onset of the colour stimuli also varied monotonically with the difference in rank. Its activity was enhanced when the to-be-attended stimulus was on the left and suppressed when it was on the right, and this differential response increased with the difference in rank (Figure 6B, top). A similar pattern also held at the population level (Figure 6B, bottom), where we used the labels obtained by k-Means clustering (Figure 5C) to assign units to the left and right-coding groups. Thus, units in the networks that are trained with the

AuGMEnT learning rule acquire a selectivity that resembles the tuning of neurons in the dorsolateral prefrontal cortex of monkeys.

In the next section we will investigate one important remaining question about the solution that was learned by these neural networks. How do networks deal with the motion stimuli, and in particular, how do they filter out stimuli on the distracter side, while monitoring stimuli on the target side as they instruct the model to release the button?

Response to the motion stimulus on the relevant and irrelevant side

To illustrate the attentional filtering mechanism, we will first focus our analysis on the example 'left' coding unit that has been illustrated in Figures 4, 5A and 6B (indexed by a light blue circle). We recorded the unit's activation in response to the first motion stimulus for three colour pairs of increasing distance. We compared two types of trials; 'respond' trials where the target stimulus was on the left side and the motion changed occurred on the same side, and 'ignore' trials where the target stimulus was on the right but the first motion change occurred on the left. These two trials types are of interest, because the motion stimulus occurs on the left, but in one case it is a target and in the other case it is a distractor, so we can specifically study the effect of attentional filtering. Figure 7 shows the influence of the left motion stimulus on the activity of the association unit when attended (left panel) versus when it needs to be ignored (right panel). When the left stimulus needs to be attended, the inputs from the colour input units brought the activity close to the steep part of the unit's non-linear activation function so that the additional motion input causes a substantial increase in activity. This increased activity enhanced the Q-value of the release action because there was an

excitatory connection from this memory unit onto the Q-value unit coding the release-action.

In contrast, when the left stimulus had to be ignored, the unit received less input from the colour input units so that its activity was farther from the steep part of the activation function. Now the motion input caused a smaller increase in activity so that it is effectively ignored because it did not lead to a large increase of the Q-value of the release action. Most networks that we investigated employed this mechanism but we also found the inverse solution, where the activity of the memory unit was high and the motion stimulus on the relevant side inhibited the memory unit, which in turn disinhibited the 'release' action through an inhibitory connection. It is intriguing that this attentional filtering problem can be solved by an appropriate weighting of colour inputs to memory units, without an influence of feedback connections on the firing rate in sensory modules (which was absent in our model; see the next section). In our model, the feedback connections are only necessary for guiding plasticity. Our results obviously do not exclude that top-down effects on firing rates in sensory cortices are essential in other tasks (such as visual search). Our analysis thereby provides insight in how attentional filtering can be implemented by a simple feedforward neuronal network, and how it can be learned with a biologically plausible reinforcement learning scheme.

Learning of feedback connections

Finally, we investigated how the amount of top-down attention for the pattern on the left and right side evolve during learning. Although in the current model the feedback connections from the response selection stage to the association units are only used to gate plasticity and do not influence the activity of units at earlier processing levels, our main aim was to investigate how these feedback connections can be learned. In order to

assess how feedback connections change during the course of learning, we measured the summed feedbacks arriving at memory units for the different trial types throughout learning for the 100 trained networks. We compared the amount of feedback from the response selection stage (i.e. through the connections marked by dashed lines in Figure 1B) to 'left' and 'right' memory units (Figure 5C) in trials where attention had to be directed to the left and to the right and integrated the total amount of feedback that arrived during the trials (Figure 8). Specifically, for each trial of duration T that networks experienced, we computed the quantity $\sum_{0<t\leq T} fb_m^M(t)$ (see equation (10)) for each memory unit.

The effect of learning is clear – during initial learning, when weights are random, units tend to receive an equal and increasing amount of feedback regardless of the trial type. After about 10% of the training time, the strength of the feedback to the units coding for the distractor side started to decrease, whereas the amount of feedback to the target side increased slightly until the end of training. Thus, during the learning process, units that code information about the target receive more feedback from the response selection stage than units that code information about the distracter.

Discussion

In recent years important studies have started to document how rewards teach attention in human and non-human primates. When subjects receive a high reward for a particular stimulus, then this stimulus is likely to attract more attention at a later point in time (B. A. Anderson et al., 2011; Chelazzi et al., 2013; Hickey et al., 2010; Libera & Chelazzi, 2009; Raymond & O'Brien, 2009). As a general finding, stimuli that have been associated with a high reward evoke stronger neuronal responses than stimuli that have been associated with lower rewards in many brain structures including the motor cortex (Pastor-Bernier & Cisek, 2011), the parietal cortex (Peck, Jangraw, Suzuki,

Efem, & Gottlieb, 2009; Platt & Glimcher, 1999), and visual cortex (Serences, 2008; Stănişor et al., 2013). In primary visual cortex, the neurons that are influenced by visual attention are also the ones that are influenced by reward expectancy (Stănişor et al., 2013), which suggests that there is a unified selection mechanism that is driven by reward expectancy as well as by shifts of attention (Maunsell, 2004). Such an effect of the reward contingency on the distribution of attention is expected because the contingency determines which information is task-relevant and which information can be ignored. In other words, there are strong theoretical grounds to believe that reward indeed controls attention. However, the precise mechanisms that explain how a reward in one trial can influence the deployment of attention in a later trial have not been well understood.

Here we have shown how a neural network trained with a biologically plausible reinforcement learning rule can learn an attentional control task when the only feedback from the environment is the occasional reward for correct performance. The performance of the networks trained with AuGMEnT exhibited a number of similarities with the performance of monkeys, and the activities of network units provide new insights into the changes in neuronal tuning that emerge during learning. First, the models exhibited a pattern of generalization to unseen colour combinations that was remarkably similar to the generalization performance of the monkeys. Second, model units acquired a strong tuning to the difference in rank-order between the two colour stimuli, just as was found in the prefrontal cortex of monkeys that had been trained on the task. Third, these newly formed representations explain why pairs of stimuli with closer ranks are associated with more errors than pairs of stimuli with larger differences in rank, because the activity of units in the association layer is more similar for stimuli with closer ranks. Fourth, the simulation provided insight in how a neural network can

decrease its sensitivity to stimuli that are task-irrelevant thereby filtering out distracting information. Finally, the present results illustrate how reward 'teaches' attention. The modified representations increased the amount of attentional feedback that was sent to units coding for the relevant motion stimulus, as instructed by the colour cues.

The AuGMEnT learning rule uses two factors to gate neuronal plasticity, and their joint action at the synapse can provide a learning rule that is as powerful as the biologically implausible error-backpropagation rule (Roelfsema & van Ooyen, 2005). The first factor is a reward prediction error, which can be computed based on the difference in activity of the Q-value units that are selected in consecutive time-steps and has been included in many previous models on reinforcement learning (e.g. Ashby, Ennis, & Spiering, 2007; Dayan & Yu, 2002; Sutton & Barto, 1998). This globally released signal informs all the synapses of the network whether the outcome of the previous action was better or worse than expected. Previous neurophysiological studies have demonstrated that many dopamine neurons in the substantial nigra and ventral tegmental area carry such reward-prediction errors (Schultz, 2002). These dopamine neurons have relatively widespread connections so that many synapses in the brain could pick up this reward prediction signal, although other neuromodulatory systems such as acetylcholine (Kilgard & Merzenich, 1998) or serotonin (Liu et al., 2014) could play equivalent roles.

The second factor that gates learning is the attentional feedback from the response selection stage (Roelfsema & van Ooyen, 2005; Rombouts et al., 2012; in press). This feedback signal originates from the units that code for the action that was selected and it assigns credit to units at earlier processing levels that were responsible for this choice. The reciprocity of feedforward and feedback connections ensures that the units at the lower levels that provide the strongest input to the selected

action are also the ones to receive a strong feedback signal. They are the ones to change their synaptic strengths (they will have the tag) as instructed by the reward prediction error (the globally released neuromodulator). A role of attentional feedback in the gating of learning is supported by studies demonstrating that subjects learn more readily about attended than non-attended features and objects (Ahissar & Hochstein, 1993) (Jiang & Chun, 2001; Trabasso & Bower, 1968). Furthermore, studies in eye movement research have firmly established that attention is invariably directed to those items that that are selected for a motor response (Deubel & Schneider, 1996); (Kowler, Anderson, Dosher, & Blaser, 1995), in accordance with the proposed feedback scheme.

At first sight, our reasoning that the feedback connections gate learning and that they themselves are learned at the same time may seem circular. How can connections gate their own plasticity? The key observation is that the feedback pathways tag those synapses of the feedforward pathways that were responsible for the selected action. These tags are a prerequisite for synaptic change based on the globally released neuromodulator. If the action resulted in an outcome that was better than expected, the tagged synapses increase in strength to promote the future selection of the same action. If the outcome of the action was disappointing, then these synapses decrease in strength. The resulting improvements in the feedforward pathways need to be accompanied by equivalent changes in the feedback pathways, as the reciprocity ensures that the credit in later trials will also assigned accurately, in spite of the modified feedforward connections. In the present work, the only influence of the attentional feedback signal is the deposit of synaptic tags for credit assignment and we did not model the well-established influence of feedback connections on the firing rates in lower level brain regions (Reynolds & Chelazzi, 2004; Roelfsema, 2006). Future models that include

top-down effects on firing rates may further expand the capabilities of neural networks that are trained with AuGMEnT-like learning rules.

The present study provides new insights in how rewards can teach attention. It is remarkable that a simple network that starts with a random connectivity can learn a relatively complex task where the rank of two colour cues determines which of two stimuli needs to be monitored for a change in motion direction, by trial and error. Trial and error learning with the AuGMEnT learning rule is versatile, because the same network and learning rule can teach networks to perform different tasks, including ones that require storage of information in working memory, non-linear mappings of sensory stimuli onto motor responses and tasks that require the integration of stochastic sensory evidence for a decision (Rombouts et al., 2012; in press).

We illustrated how the reward-prediction errors of RL theory can also provide powerful learning rules for the shaping of attentional feedback connections. These new results thereby provide insight in how a perceptual system may learn to focus attention on those features that are important to solve a cognitive task. We hypothesize that similar mechanisms are at work when we learn in less constrained environments as is the case, for example, when learning to play tennis. We anticipate that future work will address the possible generalizations of AuGMEnT-like learning rules to situations that are even more challenging or that they will point out their limitations.

Acknowledgements

The work was supported by an NWO-EW grant (n. 612.066.826), an NWO-VICI grant, a Brain and Cognition grant (n. 433-09-208), the European Union Seventh Framework Program (project 269921 “BrainScaleS” and PITN-GA-2011-290011 “ABC”) and an ERC advanced Grant (n. 339490).

Figure legends

Figure 1. Task and Model. (A) When the monkeys attained fixation, two grey moving stimuli appeared. After a delay the stimuli were coloured and these colours determined which of the two stimuli had to be monitored for a motion change. After a motion change in the relevant stimulus, the monkey had to release a button whereas motion changes at the irrelevant side had to be ignored. (B) The neural network was composed of an input layer with sustained, on and off units, an association layer with regular and memory units and a motor layer with units coding for action values (Q-values). Motor units have feedback connections (dashed) to the association layer. (C) Sensory stimuli give rise to activations in the top layer of the network. Synaptic traces (blue lines) are formed on synapses that have feedforward activations (left). Then a stochastic WTA process in the motor layer selects one of the actions for execution (middle). The selected action unit informs the rest of the network that is was selected via feedback connections (dashed lines in the middle panel). The interaction of feedback signals from the selected action and feedforward signals arising from the sensory stimuli give rise to the formation of synaptic tags (orange hexagons). These tags label the synapses in the network that were responsible for the selected action (middle). Finally, after the actually executing the action selected at time t, observing a reward and selecting a new action for execution, the difference between the predicted value at time t and the value of the actually observed transition gives rise to a prediction error δ, which is encoded by a globally released neuromodulator (green cloud). Synaptic weights in the network are

updated by a simple multiplicative interaction of the tag strength and the prediction error signal (right).

Figure 2. Generalization of the model to new colour combinations. The box plots illustrate the number of errors made, on average, by 100 networks trained on the generalization version of the colour task. First, the models were trained on three colour pairs; green-blue, yellow-green and yellow-blue. Then, models were exposed to red as a novel colour, which initially only occurred in combination with the known colour yellow. After reaching criterion performance (methods), transfer learning was tested by training the model on the remaining colour pairs, red-green and red-blue. It can be seen that few additional errors were made for these new colour pairs. This scheme was repeated for the orange colour. The error pattern for a monkey trained using the same scheme is overlaid (white circles). The boxes illustrate the lower quartile, median (thick) and upper quartile, whereas the whiskers extend to most extreme data point within 1.5 multiples of the inner quartile range). Note the discontinuity of the y-axis.

Figure 3. Effect of difference in colour rank on the error rate. (A) Hit rate of monkeys Ra and Se after 3-5 months of training as a function of distance between colour ranks. Error bars denote s.e.m. (B) Average hit rates of models (N=100) throughout learning the full colour rank task as a function of distance between colour ranks. Error bars show s.e.m. The networks were trained until they reached an accuracy of 85% (see methods), which explains why the total performance is around 80% when averaged over the whole training period. Note that the scale of panels A and B differs.

Figure 4. Example activity traces of a network trained on the full colour-ranking task. Top panels show the behaviour of two memory units with task-relevant tuning. The top unit (light blue circle) responded most strongly to if the highest-ranking colour was on the left (green had a higher rank than orange), and the middle unit (grey cross) prefers the highest rank on the right. The insets at the top show the time point when the fixation marker (F, black line) and the grey stimuli appeared on the screen, and when the colours turned on (C, bottom: left stimulus colour, middle: right stimulus colour). The black vertical line indicates the onset of a motion stimulus (M). In "respond" trials the first motion change occurred on the target side, and on "ignore" trials the first change occurred on the distracter side (M1) and then on the target side (M2). The bottom panels show action values that the model predicted for the "Hold" (blue) and "Release" (red) actions.

Figure 5. Tuning of memory units to the difference in rank between the two colours. (A) Weights from "on" sensory units to example memory units (same as those shown in figure 5), ordered by rank. The solid line marks the synaptic weights from input units coding colours of different ranks in the left visual field, and the dashed line marks the weights coming from the right visual field. (B) Sum of activations due to the presence of all possible combinations of colour stimuli. (C) Scatter-plot of linear-regression parameters (see main text). These results are based on synapses between the On-cells in the input layer and the memory units in the association layer. Off-cells and regular units usually did not have strong weights. The two example neurons from panel A (and Figure 4) have been marked with a light blue circle and a grey cross. Neurons that prefer stimuli on the right have been coloured red, and neurons that prefer stimuli on the left are shown in green. Neurons that do not have strong opposite tuning for colours on

the left and right are marked in black. Colour labels were obtained by k-Means clustering (see Results).

Figure 6. Rank difference coding. (A) Rank difference coding of a single cell in the prefrontal cortex of a monkey (top) and a population of prefrontal neurons (bottom, both adapted from (Lennert & Martinez-Trujillo, 2011)). Solid lines: target pattern on the preferred side, dashed lines: target on non-preferred side. (B) Rank difference coding in a single model unit (top) and over the whole population of trained model units. Solid lines: target at preferred location, dashed lines: distracter at preferred location. Shadings show s.e.m.

Figure 7. Attentional filtering by an example 'left' unit. (A) Colour combination with the highest rank on the left brings the unit close to the steep part of the non-linear activation function (black sigmoidal curve). The appearance of the motion stimulus can therefore cause a large increase in the unit's activity. The three arrows show the increase in activity for colour combinations with a distance in rank of 1, 2 and 3. This increase in activity is propagated to the output layer to cause an increase of the Q-value of the button release action. (B) Colour combinations cueing that the right motion stimulus is relevant cause suppression so that the unit is relatively insensitive to a change in motion direction on the left side. Thus, a motion change on the irrelevant side cannot cause a strong increase in the Q-value of the release action and will be ignored by the model.

Figure 8. Learning shapes attentional feedback from the response selection stage. Mean summed attentional feedback arriving at memory units encoding information

about the target side (green) versus the distracter side (red) throughout learning. Shading shows s.e.m. We averaged feedback across all trial types. Because they did not occur equally often (due to the random generation of trial types), we estimated the relative position of trials of every type in the learning sequence with linear interpolation before averaging them.

References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(12), 5718–5722.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–10371. doi:10.1073/pnas.1104047108

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632–656. doi:10.1037/0033-295X.114.3.632

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. *Neuron*, *68*(5), 815–834. doi:10.1016/j.neuron.2010.11.022

Bundesen, C., Habekost, T., & Kyllingsbæk, S. (2005). A Neural Theory of Visual Attention: Bridging Cognition and Neurophysiology. *Psychological Review*, *112*(2), 291–328. doi:10.1037/0033-295X.112.2.291

Chelazzi, L., Perlato, A., Santandrea, E., & Libera, Della, C. (2013). Rewards teach visual selective attention. *Vision Research*, *85*, 58–72. doi:10.1016/j.visres.2012.12.005

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. doi:10.1038/nrn755

Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and Reinforcement Learning. *Neuron*, *38*, 285–298.

Dayan, P., & Yu, A. J. (2002). ACh, uncertainty, and cortical inference (Vol. 1, p. 189). Presented at the Advances in Neural Information Processing Systems, MIT Press.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, *21*(8), 355–361.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision*

Research, *36*(12), 1827–1837.
Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. doi:10.1016/j.tics.2010.01.004
Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.
Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349.
Gnadt, J. W., & Andersen, R. A. (1988). Memory related motor planning activity in posterior parietal cortex of macaque. *Experimental Brain Research*, *70*(1), 216–220.
Gottlieb, J., & Balan, P. (2010). Attention as a decision in information space. *Trends in Cognitive Sciences*, *14*(6), 240–248. doi:10.1016/j.tics.2010.03.001
Hickey, C., Chelazzi, L., & Theeuwes, J. (2010). Reward Changes Salience in Human Vision via the Anterior Cingulate. *Journal of Neuroscience*, *30*(33), 11096–11103. doi:10.1523/JNEUROSCI.1026-10.2010
Jiang, Y., & Chun, M. M. (2001). Selective attention modulates implicit learning. *The Quarterly Journal of Experimental Psychology A*, *54*(4), 1105–1124. doi:10.1080/02724980042000516
Kilgard, M. P., & Merzenich, M. M. (1998). Cortical Map Reorganization Enabled by Nucleus Basalis Activity. *Science*, *279*(5357), 1714–1718. doi:10.1126/science.279.5357.1714
Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, *35*(13), 1897–1916.
Krueger, K. A., & Dayan, P. (2009). Flexible shaping: How learning in small steps helps. *Cognition*, *110*(3), 380–394. doi:10.1016/j.cognition.2008.11.014
Lennert, T., & Martinez-Trujillo, J. (2011). Strength of Response Suppression to Distracter Stimuli Determines Attentional-Filtering Performance in Primate Prefrontal Neurons. *Neuron*, *70*(1), 141–152. doi:10.1016/j.neuron.2011.02.041
Libera, Della, C., & Chelazzi, L. (2009). Learning to attend and to ignore is a matter of gains and losses. *Psychological Science*, *20*(6), 778–784. doi:10.1111/j.1467-9280.2009.02360.x
Liu, Z., Zhou, J., Li, Y., Hu, F., Lu, Y., Ma, M., et al. (2014). Dorsal Raphe Neurons Signal Reward through 5-HT and Glutamate. *Neuron*, *81*(6), 1360–1374. doi:10.1016/j.neuron.2014.02.010
Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward Value-Based Gain Control: Divisive Normalization in Parietal Cortex. *Journal of Neuroscience*, *31*(29), 10627–10639. doi:10.1523/JNEUROSCI.1237-11.2011
Mao, T., Kusefoglu, D., Hooks, B. M., Huber, D., Petreanu, L., & Svoboda, K. (2011). Long-Range Neuronal Circuits Underlying the Interaction between Sensory and Motor Cortex. *Neuron*, *72*(1), 111–123. doi:10.1016/j.neuron.2011.07.029
Maunsell, J. H. R. (2004). Neuronal representations of cognitive state: reward or attention? *Trends in Cognitive Sciences*, *8*(6), 261–265.

doi:10.1016/j.tics.2004.04.003
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202.
Pastor-Bernier, A., & Cisek, P. (2011). Neural Correlates of Biased Competition in Premotor Cortex. *Journal of Neuroscience*, *31*(19), 7083–7088. doi:10.1523/JNEUROSCI.5681-10.2011
Peck, C. J., Jangraw, D. C., Suzuki, M., Efem, R., & Gottlieb, J. (2009). Reward Modulates Attention Independently of Action Value in Posterior Parietal Cortex. *Journal of Neuroscience*, *29*(36), 11182–11191. doi:10.1523/JNEUROSCI.1929-09.2009
Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238.
Raymond, J. E., & O'Brien, J. L. (2009). Selective Visual Attention and Motivation The Consequences of Value Learning in an Attentional Blink Task. *Psychological Science*, *20*(8), 981–988.
Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611–647. doi:10.1146/annurev.neuro.26.041002.131039
Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annual Review of Neuroscience*, *29*, 203–227.
Roelfsema, P. R., & van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural Computation*, *17*(10), 2176–2214.
Roelfsema, P. R., van Ooyen, A., & Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends in Cognitive Sciences*, *14*(2), 64–71. doi:10.1016/j.tics.2009.11.005
Rombouts, J. O., Bohte, S. M., & Roelfsema, P. R. (2012). Neurally Plausible Reinforcement Learning of Working Memory Tasks (Vol. 25, pp. 1880–1888). Presented at the Advances in Neural Information Processing Systems.
Rombouts, J. O., Bohte, S. M., & Roelfsema, P. R. (in press). How Attention Can Create Synaptic Tags for the Learning of Working Memories in Sequential Tasks. *PLoS Computational Biology*.
Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (No. CUED/F-INFENG/TR 166). Cambridge.
Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.
Serences, J. T. (2008). Value-Based Modulations in Human Visual Cortex. *Neuron*, *60*(6), 1169–1181. doi:10.1016/j.neuron.2008.10.051
Stănişor, L., van der Togt, C., Pennartz, C. M., & Roelfsema, P. R. (2013). A unified selection signal for attention and reward in primary visual cortex. *Proceedings of the National Academy of Sciences*, *110*(22), 9136–9141. doi:10.1073/pnas.1300117110/-/DCSupplemental
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: an introduction*. MIT Press.
Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research*. Krieger.
Treue, S., & Maunsell, J. H. (1996). Attentional modulation of visual motion

processing in cortical areas MT and MST. *Nature*, *382*(6591), 539–541. doi:10.1038/382539a0

Whitehead, S. D., & Ballard, D. H. (1991). Learning to perceive and act by trial and error. *Machine Learning*, *7*(1), 45–83.

Figure 1: Task and Model.

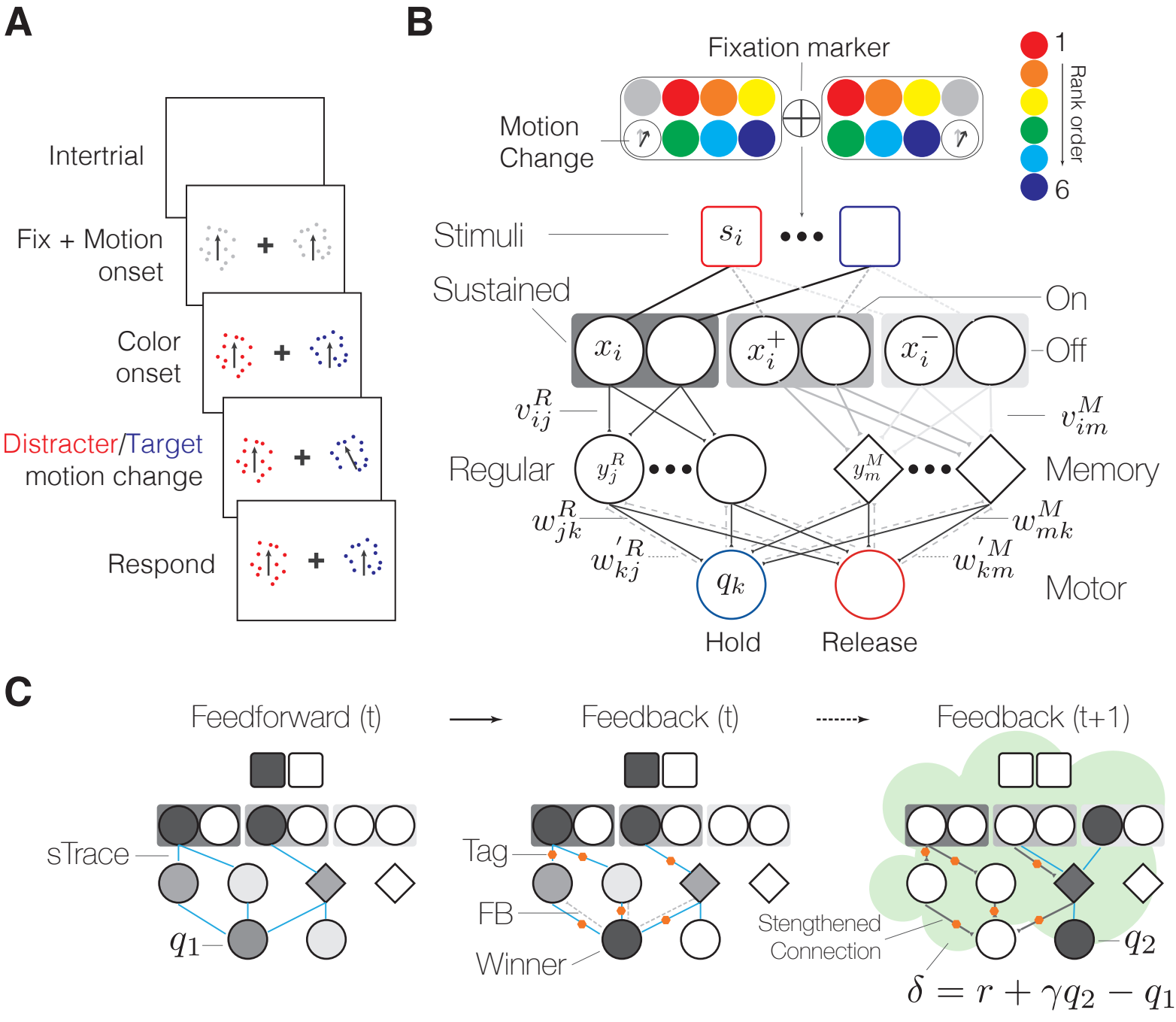

Figure 2. Generalization of the model to new colour combinations.

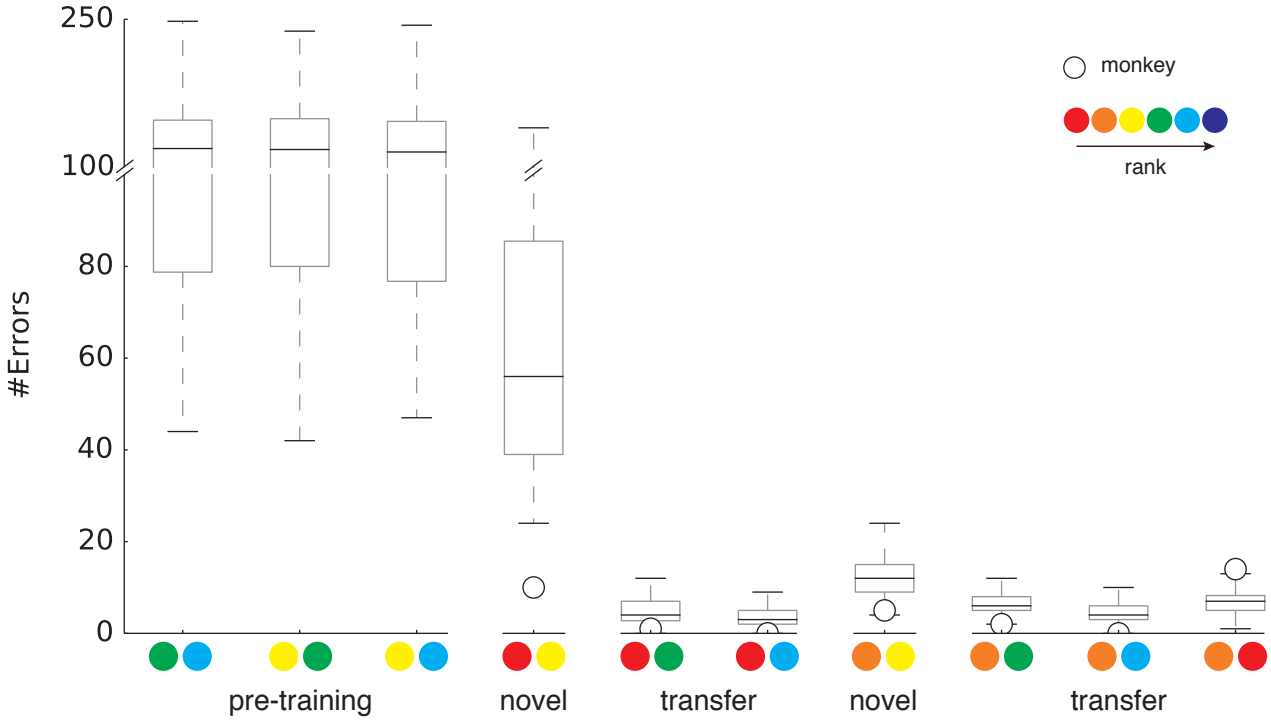

Figure 3. Effect of difference in colour rank on the error rate.

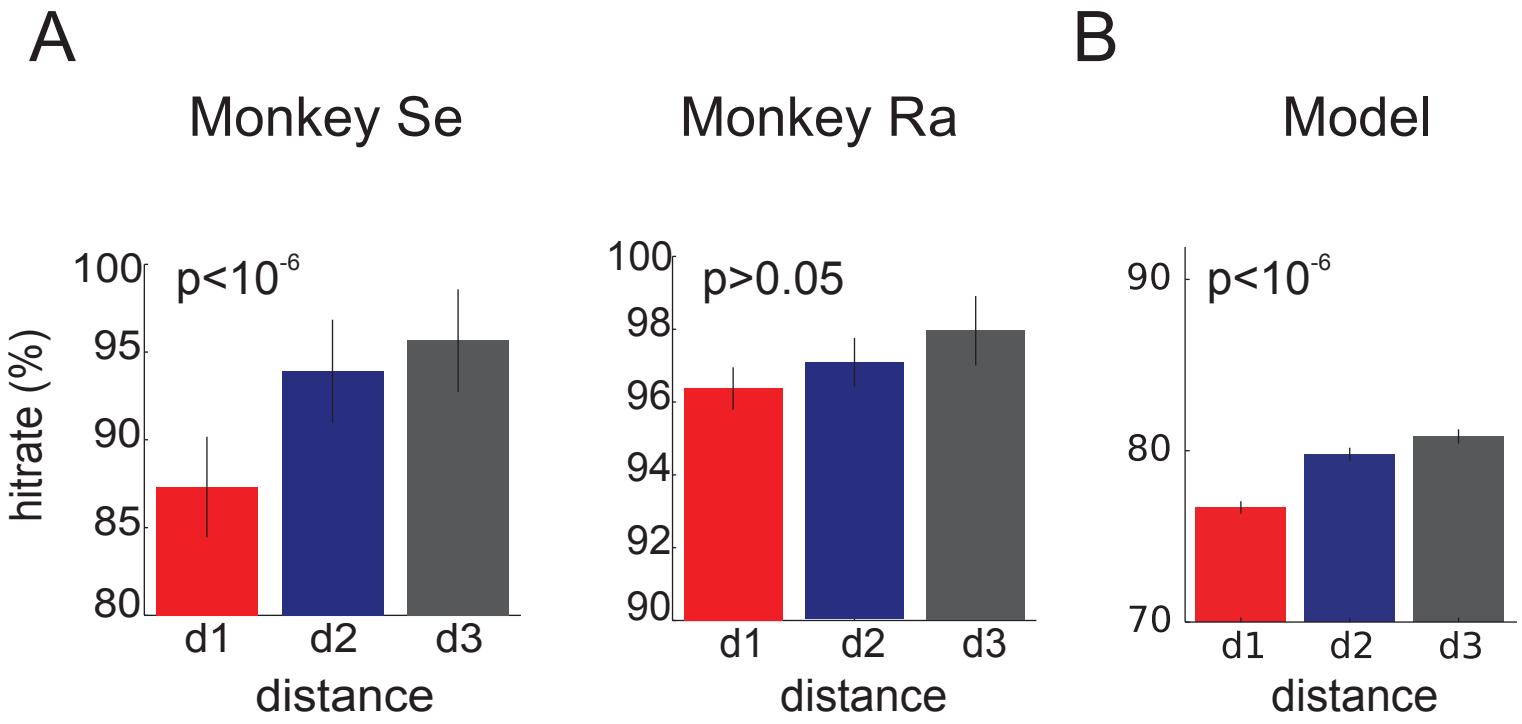

Figure 4. Example activity traces of a network trained on the full colour-ranking task.

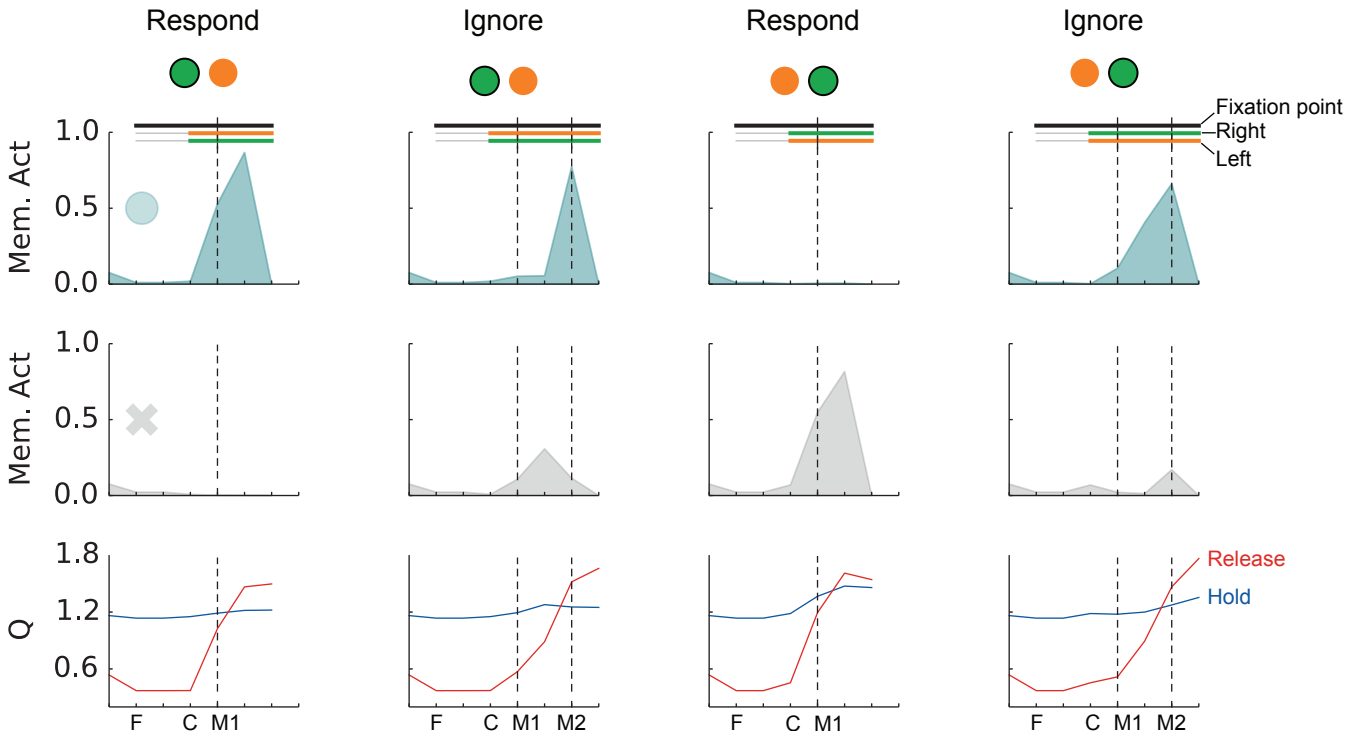

Figure 5. Tuning of memory units to the difference in rank between the two colours.

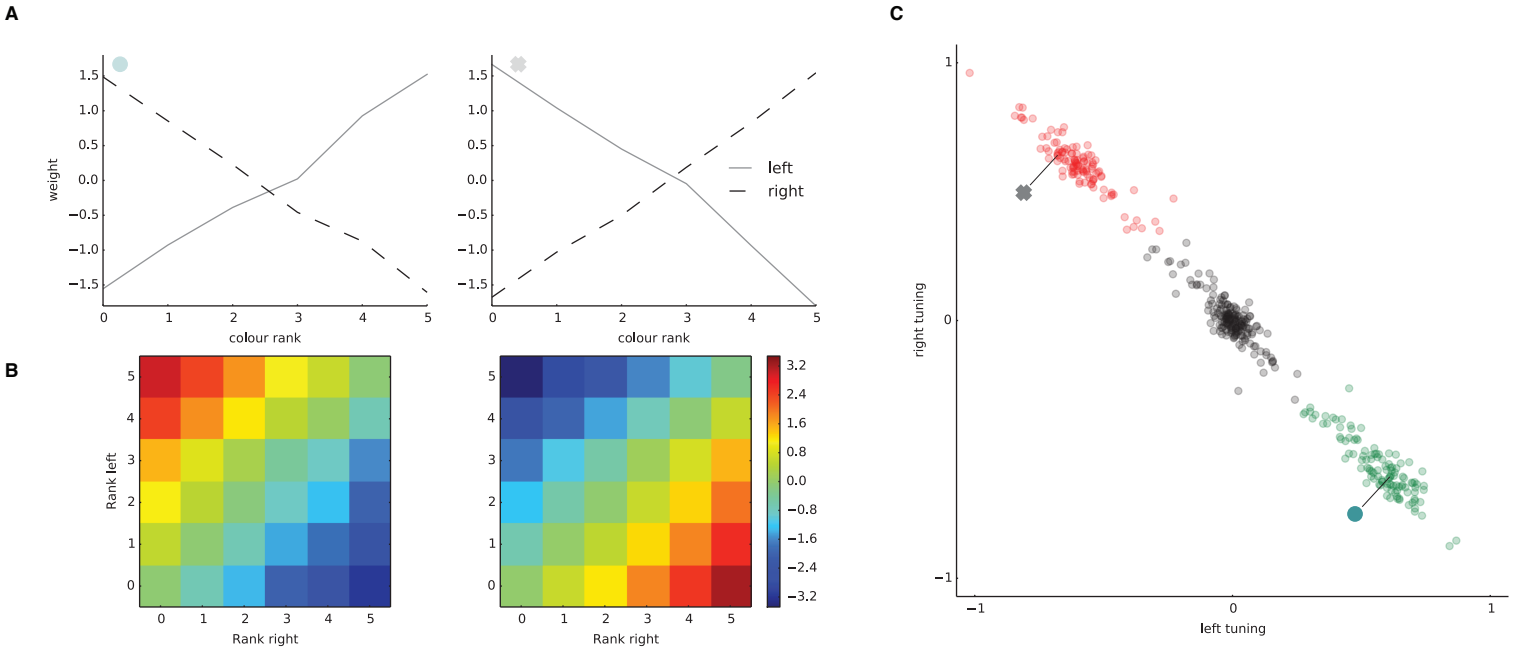

Figure 6. Rank difference coding.

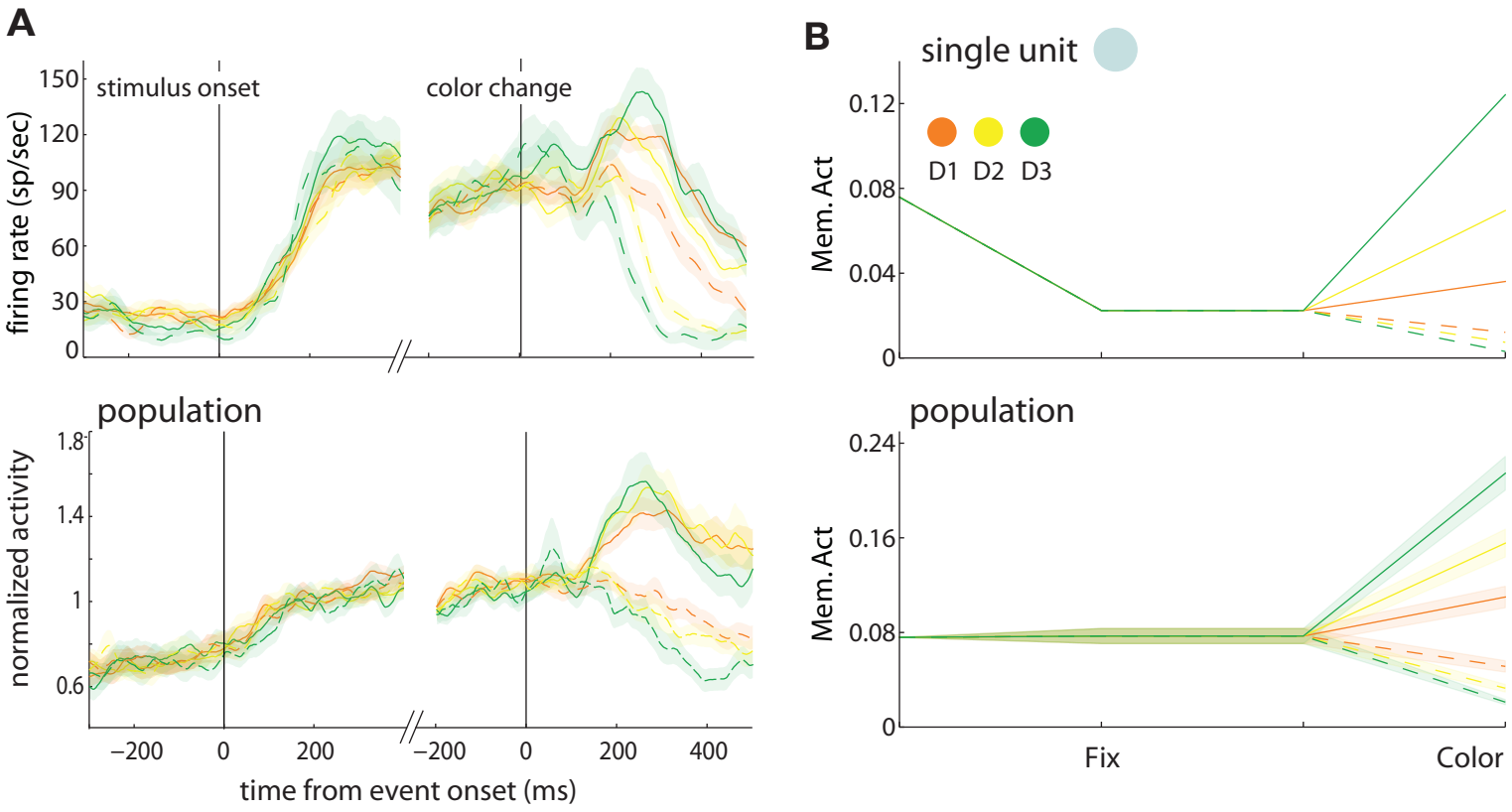

Figure 7. Attentional filtering by an example ‘left’ unit.

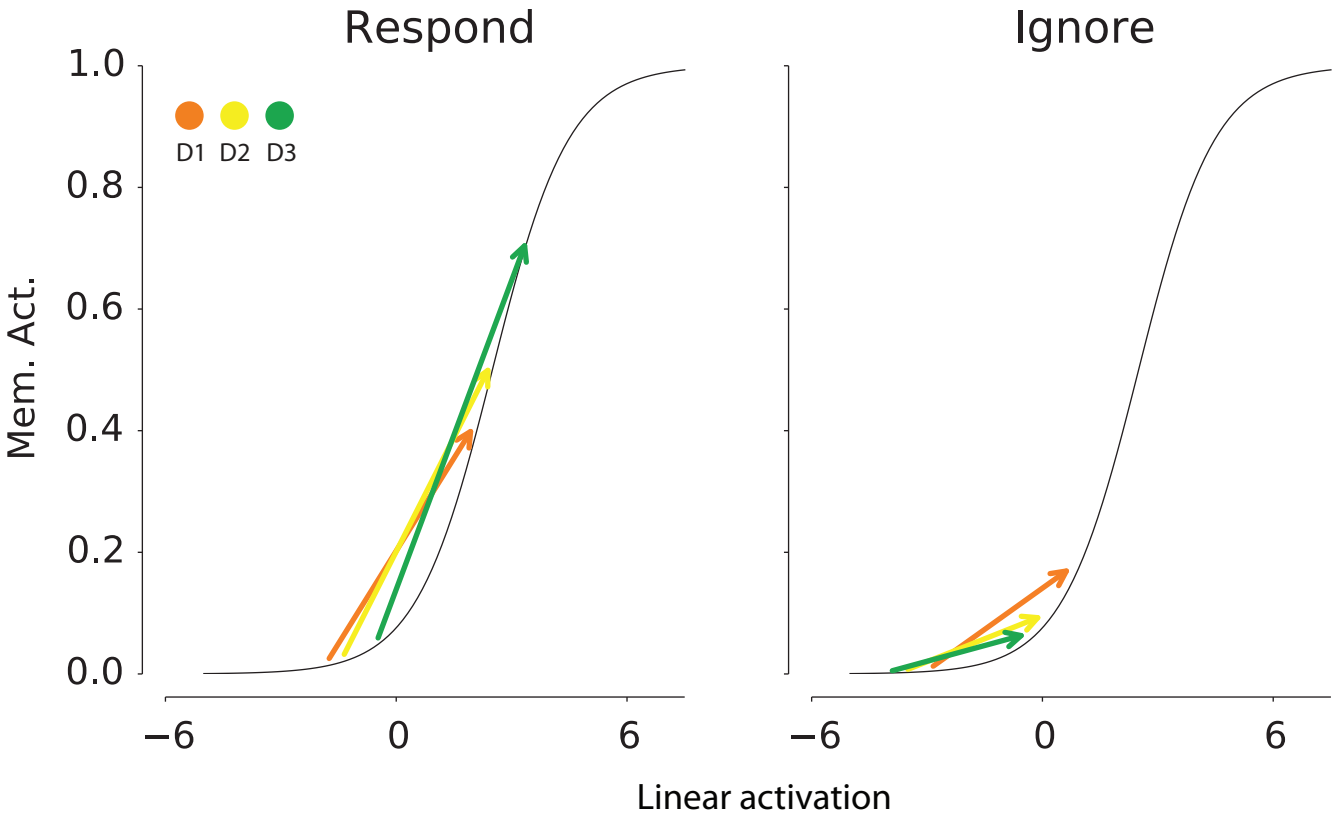

Figure 8. Learning shapes attentional feedback.