International Journal of Psychophysiology 127 (2018) 62–72

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Learning what matters: A neural explanation for the sparsity bias

Cameron D. Hassall[a,*], Patrick C. Connor[b], Thomas P. Trappenberg[b], John J. McDonald[c], Olave E. Krigolson[a]

[a] Centre for Biomedical Research, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada
[b] Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada
[c] Department of Psychology, Simon Fraser University, Vancouver, British Columbia V5A 1S6, Canada

ARTICLE INFO

Keywords:
Attention
Learning
Reward positivity
N2pc
Sparsity

ABSTRACT

The visual environment is filled with complex, multi-dimensional objects that vary in their value to an observer's current goals. When faced with multi-dimensional stimuli, humans may rely on biases to learn to select those objects that are most valuable to the task at hand. Here, we show that decision making in a complex task is guided by the sparsity bias: the focusing of attention on a subset of available features. Participants completed a gambling task in which they selected complex stimuli that varied randomly along three dimensions: shape, color, and texture. Each dimension comprised three features (e.g., color: red, green, yellow). Only one dimension was relevant in each block (e.g., color), and a randomly-chosen value ranking determined outcome probabilities (e.g., green > yellow > red). Participants were faster to respond to infrequent probe stimuli that appeared unexpectedly within stimuli that possessed a more valuable feature than to probes appearing within stimuli possessing a less valuable feature. Event-related brain potentials recorded during the task provided a neurophysiological explanation for sparsity as a learning-dependent increase in optimal attentional performance (as measured by the N2pc component of the human event-related potential) and a concomitant learning-dependent decrease in prediction errors (as measured by the feedback-elicited reward positivity). Together, our results suggest that the sparsity bias guides human reinforcement learning in complex environments.

1. Introduction

Humans are consistently faced with complex decision problems in multidimensional environments (environments involving choice stimuli made up of numerous features). For example, a decision to eat at one of several restaurants may involve many aspects, including the type of food, ambiance, speed of service, price, availability of parking, location, and existing reviews of the establishment. When combined with other factors (style of cooking, appearance, availability of parking, number of staff, location, reputation, etc.) the state space representing all possible restaurant choices is enormous, yet most humans have little difficulty deciding where to eat. In contrast, reinforcement learning (RL) algorithms that accurately predict behavior in simple tasks become bogged down when faced with multidimensional choices (Sutton and Barto, 1998; also see *curse of dimensionality*: Bellman, 1957). It is therefore problematic that although evidence suggests human learning depends in part on an RL system implemented within dopaminergic midbrain neurons (Roesch et al., 2012; Schultz, 2013) and medial frontal cortex (Holroyd and Coles, 2002; Krigolson et al., 2014; Krigolson et al., 2009; Sambrook and Goslin, 2015), we are able to solve complex multidimensional problems faster than an RL approach alone would predict.

At their core, RL algorithms compare actual feedback with expected feedback to compute prediction errors (Sutton and Barto, 1998), which are then used to update the weights associated with chosen actions. This update process ensures that, in the long run, actions are taken that are more likely to maximize utility (Mill, 1863). Although traditional RL algorithms eventually converge upon optimal solutions, they may do so slowly when compared to an ideal (Bayesian) model. For example, a traditional RL algorithm may select or avoid previously-chosen multidimensional stimuli only insofar as those exact combinations of features are re-encountered. Thus, traditional RL algorithms may be unable to learn about dimensions, only about combinations of features across all dimensions. Consider the following: positive experiences at Italian restaurants in two different areas of town should probably reinforce choosing Italian restaurants in general. However, without rewarding visits to additional Italian restaurants, traditional RL models will only reinforce choosing the two previously visited locations.

The key to the above example is that sometimes only a subset of features is predictive of reward (e.g., style of cooking). Such an

[*] Corresponding author at: School of Exercise Science, Physical and Health Education, University of Victoria, P.O. Box 17000 STN CSC, Victoria, British Columbia V8W 2Y2, Canada.
E-mail address: chassall@uvic.ca (C.D. Hassall).

https://doi.org/10.1016/j.ijpsycho.2018.03.006
Received 18 October 2017; Received in revised form 6 March 2018; Accepted 9 March 2018
Available online 15 March 2018

environment is called "sparse" and recent work suggests that humans are able to exploit this property, when it is present (Gershman et al., 2010; Leong et al., 2017; Niv et al., 2015; Wilson and Niv, 2012). For example, Gershman et al. (2010) proposed that humans use an inductive bias – a decision-making assumption – called sparsity to guide learning in multidimensional worlds. The authors presented participants with choice stimuli that varied along three dimensions (shape, color, and texture). Importantly, only one dimension was ever relevant for predicting reward. Gershman et al. (2010) compared human performance on this task with that of a Bayesian model that estimated the posteriors of each feature being the target feature (the likelihoods, given the feedback history). They observed that while humans may not choose optimally (i.e., according to Bayes' rule), they performed better than a traditional RL model (what they called "naive RL") predicted. Based on these findings, Gershman et al. (2010) proposed that humans employ a hybrid RL/Bayesian approach that guides feedback-based learning by selectively attending to a single dimension – the dimension currently believed to be most relevant.

In the present study we assumed that using the sparsity bias would engage two neural processes: an attentional selection process, and an RL process. We therefore hypothesized that a neural marker for each of these processes would be evident in a task for which the use of sparsity was beneficial. We further hypothesized that those neural markers would change with learning as predicted by an attention-weighted RL model (the hybrid RL model proposed by Gershman et al., 2010). We used two event-related potential (ERP) components as neural markers to assess the contributions of both RL and attentional systems within the brain to decision making. First, we measured the N2pc component as an index of selective attention (the proposed mechanism behind the sparsity bias: Gershman et al., 2010; Leong et al., 2017; Niv et al., 2015). Second, we used the reward positivity (Holroyd et al., 2008; Krigolson et al., 2014),[1] also known as the feedback-related negativity (FRN: Miltner et al., 1997), as a neural marker for RL.

A strong relationship between attention and learning is supported by both behavioral and ERP studies. Indeed, our experiences teach us to focus our attention on task-relevant features (Mackintosh, 1975; Pearce and Hall, 1980; Dayan et al., 2000). For example, Wills et al. (2007) showed ERP evidence that unexpected wins (i.e., positive RL prediction errors) grab our attention and drive learning. They did this using stimulus-response learning tasks involving combinations of features. Features that were not expected to elicit a reward (but did) later elicited a large N1, an ERP component associated with the allocation of visual attention to a spatial location (Hillyard and Anllo-Vento, 1998). Eye tracking data confirm that we learn the most about what we attend to; gaze duration for a cue is reduced if it is paired with a previously-rewarded stimulus, or *blocked* (Kruschke et al., 2005). Importantly, attention is not only directed to spatial locations per se, but may also be directed toward features and even feature dimensions (dimensional weighting: Müller et al., 2003).

Our neural marker of selective attention, the N2pc is a negative potential at posterior electrodes that is enhanced contralateral to attended objects appearing in multi-item arrays (Luck and Hillyard, 1994a, 1994b). While it is often observed in response to singletons (single-dimensional stimuli), the N2pc can also be elicited by targets defined by a conjunction of features (e.g., shape and color: Luck et al., 1993). Although there continues to be debate about whether the N2pc reflects enhancement of attended items or suppression of unattended items, researchers agree that the N2pc is tied to an early stage of attentional selection (Eimer, 1996; Hickey et al., 2009; Luck and Hillyard, 1994a, 1994b). Thus, learning biases such as sparsity that rely on selective attention should elicit an N2pc component in response to features that are predictive of reward. Furthermore, we would expect such a component to be dependent on learning. In particular, we might expect an enhanced N2pc component later in learning, when attention is focused on relevant features, compared to early in learning, when the identity of the relevant dimension may be unknown.

While there are currently several ways to measure RL signals in humans (Niv, 2009), a growing body of evidence suggests that the reward positivity indexes a generic RL system within the human brain (Chase et al., 2015; Holroyd and Coles, 2002). The reward positivity component is differentially sensitive to gains and losses (more positive for gains, more negative for losses) and appears 250 to 350 ms after feedback over frontal-central scalp regions. Holroyd and Coles (2002) and others have suggested that the reward positivity reflects an RL prediction error. Specifically, the reward positivity is enhanced for unexpected gains/losses compared to expected gains/losses (Holroyd and Coles, 2002; Holroyd and Krigolson, 2007; Holroyd et al., 2003; Holroyd et al., 2008; Oliveira et al., 2007). Furthermore, the amplitude of the reward positivity tends to decrease with learning as feedback becomes more expected (Krigolson et al., 2014; Krigolson et al., 2009). Finally, the reward positivity appears to shift through time to the earliest indicator that events are better or worse than expected (Holroyd et al., 2011; Krigolson and Holroyd, 2007; Krigolson et al., 2014). (See Niv, 2009, for other ways to measure RL signals in humans.) Thus, two types of reward positivity evidence were considered in the present study. First, the presence (existence) of a reward positivity would indicate the activity of an RL system sensitive to rewards and punishments. Second, changes in the amplitude of the reward positivity ought to behave as predicted by an RL model (e.g., diminish with learning).

The present study tracked changes in two ERP components – the N2pc and the reward positivity – in order to test whether or not humans use the sparsity bias and RL when choosing complex stimuli. If selective attention is engaged, we should observe a (learning dependent) enhancement of the N2pc in response to reward-predicting features, as predicted by a hybrid RL/Bayesian model. If an RL system is engaged, we should observe a reward positivity that decreases with learning, consistent with model-generated prediction errors. We tested these hypotheses by modifying a decision-making task used in previous studies to evaluate the sparsity bias (Gershman et al., 2010; Leong et al., 2017; Niv et al., 2015; Wilson and Niv, 2012). Specifically, we had participants select one of two multidimensional stimuli in which only a single dimension was predictive of reward, and we examined the ERPs elicited by the choice stimuli (for evidence of an attentional bias) and by the subsequent feedback (for evidence of RL), thus providing the first ERP evidence for the sparsity bias.

2. Experimental procedures

2.1. Participants

We tested 16 participants with no known neurological impairments and with normal or corrected-to-normal vision (4 male, μ_{age} = 18.8 years, σ_{age} = 1.1 years). All participants provided informed consent approved by the Health Sciences Research Ethics Board at Dalhousie University, and the study was conducted in accordance with the ethical standards prescribed in the original (1964) and subsequent revisions of the Declaration of Helsinki.

2.2. Apparatus and procedure

Participants were seated 75 cm in front of a 22-inch LCD monitor (75 Hz, 2 ms response rate, 1680 by 1050 pixels, LG W2242TQ-GF, Seoul, South Korea). Visual stimuli were presented using the Psychophysics Toolbox Extension (Brainard, 1997; Pelli, 1997) for MATLAB (Version 8.2, Mathworks, Natick, USA). Participants were given both verbal and written instructions in which they were asked to minimize head and eye movements.

Participants completed a decision-making task of 40 blocks of 20

[1] For simplicity we will from this point use the term reward positivity. See Proudfit (2015) for a discussion on the definition and naming of the reward positivity/FRN.

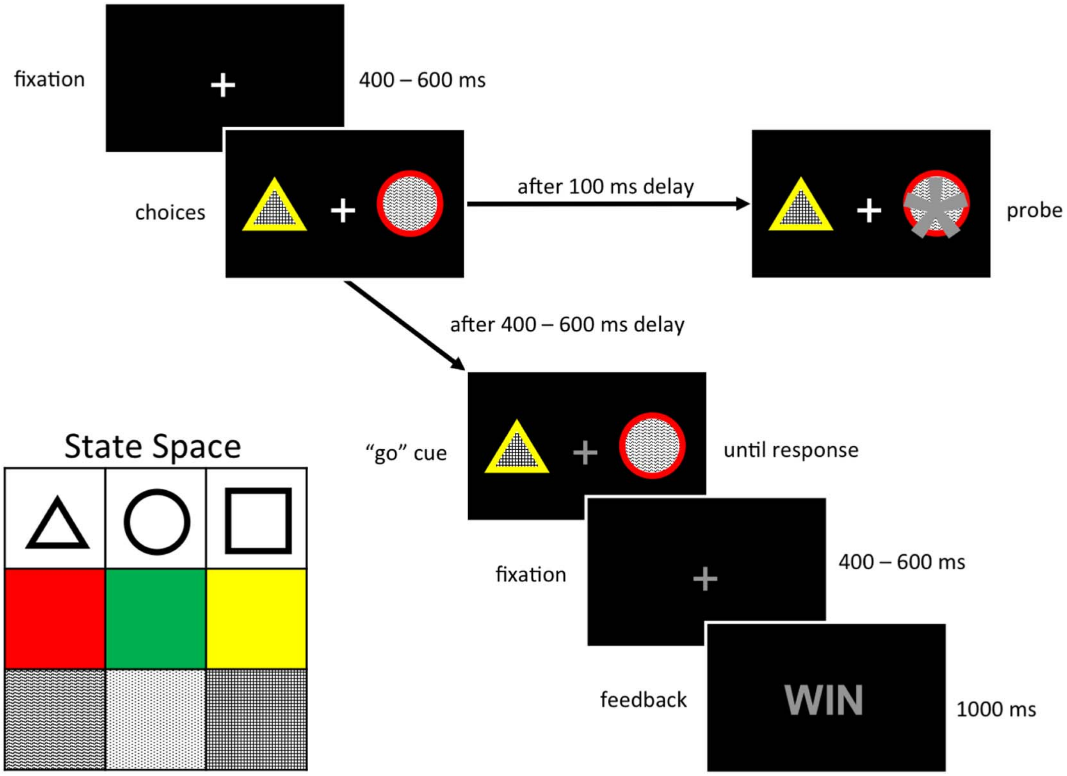

Fig. 1. Experimental task, with timing details. Choice stimuli were randomly generated for each trial by sampling from a state space with three dimensions (shape, color, and texture) of three properties each. On randomly chosen trials, a star probe would appear shortly after stimulus presentation.

trials each. Participants were told that the goal of the task was to win as many points as possible. Trials began with a white fixation cross on a black background. After 400–600 ms, two figures appeared on either side of the fixation cross. Figures varied along three dimensions: shape, color, and texture. Participants were informed that on each trial the figures would randomly vary along these dimensions by sampling from the following sets, without replacement: shape (square, circle, or triangle), color (red, green, or yellow), and texture (crosshatch, waves, or dots). See Fig. 1 for sample stimuli. Participants were instructed to select one of the figures by pressing either the left or right button on a USB gamepad in order to select either the left or right figure, respectively. Following a short delay (400–600 ms), participants were given feedback for their choice – either the word "WIN" or "LOSE". Wins resulted in a gain of two points, and losses resulted in a loss of one point. Participants were shown their point total at the end of the experiment.

Participants were also told that on each trial selecting one of the figures was more likely to lead to a win than selecting the other. In particular, it was explained that each block of 20 trials had only a single target dimension (shape, color, or texture) – that is, only one dimension was relevant for predicting outcomes within that block. The target dimension – either shape, color, or texture – was randomly chosen at the beginning of each block. The identity of the target dimension was never revealed to participants. It was further explained that within a block's target dimension there was a ranking of property values. For example, if shape was the relevant dimension within a block, the value ranking might be "square, circle, triangle" so that square stimuli are more likely to yield rewards than circle stimuli, and circle stimuli are more likely to yield rewards than triangle stimuli. Participants were told that the value rankings determined outcome likelihoods; selecting the figure with the highest-valued property would usually lead to a win, selecting the figure with the middle-valued property would sometimes lead to a win, and selecting the figure with the lowest-valued property would almost never result in a win but a loss. The precise outcome likelihoods were not revealed to participants (0.95 for the highest-valued, 0.5 for the middle-valued, and 0.05 for the lowest-valued).

While the target dimension within each block could take on one of three ranked properties, only two of these properties were shown on each trial since only two figures appeared on the display. Thus, while the highest-valued property may not have been present on every trial, there was always a higher-valued option (e.g., the figure with the middle-valued property, if the alternative had the lowest-valued property). Participants were advised to do their best in selecting the figure most likely to lead to a win. Note that in the original version of this task (Gershman et al., 2010) participants chose from among three figures – all properties were present at all times. Here, we restricted the choice to two figures in order to measure the N2pc, which relies on differential processing between images in the left visual field and images in the right visual field.

To measure the behavioral impact of value on attention, on several randomly selected trials (called "catch trials") an attentional probe would appear on either the left or the right side of the display (a dot probe: MacLeod et al., 1986). The purpose of including the probe was to elicit a behavioral measure of attention alongside our ERP measure of attention (the N2pc). The probe appeared as a star within one of the figures, shortly after choice presentation (around 100 ms after the figures appeared – see Fig. 1). Participants were instructed to respond to the probe by pressing any button on the gamepad as quickly as possible. Catch trials occurred every four, five, or six trials (chosen randomly). In an effort to maintain participants' focus on the main decision-making task, probes never appeared on trials one through four within a block (the first probe thus appeared on trial five, six, or seven).

2.3. Data collection

Our experimental software recorded reaction times for catch trials (time since onset of star probe) as well as the ranking of the target dimension property for the figure that the probe appeared within (high, medium, or low). For all other trials, participant choices were recorded, as well as the corresponding target dimension property (high, medium, or low). Additionally, the "correctness" of each choice was recorded, that is, whether or not the higher-valued stimulus was chosen.

EEG was recorded from 64 electrode locations using Brain Vision PyCorder software (Version 1.0.4, Brain Products GmbH, Munich, Germany). The electrodes were mounted in a fitted cap with a standard 10–20 layout and were recorded with respect to a virtual ground built into the amplifier (i.e., reference-free acquisition). The vertical and horizontal electrooculograms were recorded from electrodes placed above and below the right eye and on the outer canthi of the left and right eyes. Electrode impedances were below 20 kΩ before recording began and the EEG data were sampled at 500 Hz and amplified (ActiCHamp, Brainproducts GmbH, Munich, Germany).

2.4. Data analysis

2.4.1. Behavioral

Catch trials with reaction times < 200 ms or > 2000 ms were removed from subsequent data analysis. Catch trial reaction time means and standard deviations were computed for each participant and condition (higher-valued catch trials and lower-valued catch trials). To reiterate: the higher-valued stimulus was defined as the stimulus most likely to yield a win when selected. Accuracy, defined as the proportion of times across all blocks that the higher-valued stimulus was chosen, was computed for each trial and participant. We also computed accuracy mean and standard deviation across all participants for each trial. Finally, we computed decision time means and standard deviations on non-catch trials, grouped by stimulus choice (high/low). Recall that choosing the higher-valued stimulus was considered correct, and choosing the lower-valued stimulus was considered incorrect. In order to mirror our ERP analysis, accuracies, decision times, and catch trial reaction times were further grouped into early trials $(1-10)$ and late trials $(11-20)$. Note that the distribution of catch trials within each block resulted in more late catch trials than early catch trials. As a validity check, we repeated our catch trial analysis using only the first and last catch trial within each block.

2.4.2. EEG

EEG data were first downsampled to 250 Hz and rereferenced to the average of the two mastoid channels. The data were then filtered through a (0.1 Hz – 30 Hz pass band, 60 Hz notch) phase shift-free Butterworth filter. Next, 800 ms epochs of data were extracted, beginning 200 ms before our events of interest (the onset of the choice stimuli and the onset of feedback). Next, ocular artifacts were removed using independent component analysis. Subsequent to this, all trials were baseline corrected using a 200 ms epoch prior to stimulus onset. Finally, trials in which the change in voltage in any channel exceeded 10 μV per millisecond or the change in voltage across the epoch was > 100 μV were discarded. In total, < 10% of the data were discarded. See Table 1 for the mean number of trials that were recorded for each condition, and the total number of trials that survived artifact rejection. We analyzed responses to two events: the N2pc, locked to the appearance of the choice stimuli, and the reward positivity, locked to the appearance of feedback.

2.4.2.1. N2pc. To evaluate responses to the appearance of the choice stimuli, 800 ms epochs of data (from 200 ms before choice onset to 600 ms after choice onset) were extracted from the continuous EEG for all trials. As with our behavioral analysis, we grouped these epochs based on trial number (early: trials 1–10 and late: trials 11–20) and whether the higher-valued choice appeared on the left or right side of the display ("high-left" or "high-right" trials). In order to determine the relationship between the N2pc and behavior, early epochs (trials 1–10) were also grouped based on participant choice: correct (if the higher-valued stimulus was chosen) and incorrect (if the lower-valued stimulus was chosen). Similar epochs were analyzed for the response to feedback (from 200 ms before to 600 ms after feedback onset). Feedback epochs were further grouped by trial as before (early/late), and by feedback type ("win" or "loss" trials).

Next, we created ERPs for each participant by averaging the stimulus-locked EEG data at each electrode channel for all high-left and high-right trials, both early and late in learning (early high-left, early high-right, late high-left, and late high-right). Consistent with previous work on the N2pc (e.g., Kiss et al., 2008), a second average was computed by combining the EEG signals from electrode channels PO7 and PO8, depending on whether that electrode site was contralateral or ipsilateral to the higher-valued choice. Both high-left and high-right trials thus contributed to each of the contralateral and ipsilateral averages and the process of averaging the signals from PO7 and PO8 created four new data sets: early/late contralateral PO7/8 and early/late ipsilateral PO7/8. The N2pc component was analyzed using the difference wave method: the average ipsilateral waveform was subtracted from the average contralateral waveform both early and late in learning. The N2pc component was then defined as the mean voltage in the average PO7/8 difference waveforms 220–300 ms post stimulus, based on the location of the peaks of the grand average waveform and on previous work (Eimer and Kiss, 2008; McDonald et al., 2013).

2.4.2.2. Reward positivity. Similarly, the EEG response to feedback was quantified by averaging EEG data for each participant, channel, and feedback condition (win/loss) both early and late in learning. The reward positivity was then analyzed using the difference wave method: the average waveform for loss trials was subtracted from the average waveform for win trials for each participant both early and late in learning. Previous work (Holroyd and Coles, 2002; Holroyd and Krigolson, 2007; Krigolson et al., 2014; Krigolson et al., 2009; Miltner et al., 1997) and an examination of the grand average difference wave (all conditions combined, not shown) led us to define the reward positivity as the mean deflection of the difference waveform

Table 1
Mean number of recorded trials and trials remaining after artifact rejection (AR).

	Recorded		Remaining after AR	
	Trial count	95% CI	Trial count	95% CI
Early win (trials 1–10)	201	[188, 213]	187	[171, 202]
Late win (trials 11–20)	226	[211, 241]	213	[193, 232]
Early loss (trials 1–10)	177	[168, 186]	163	[153, 174]
Late loss (trials 11–20)	154	[142, 166]	142	[131, 153]
Early high-left (trials 1–10)	202	[195, 209]	186	[174, 198]
Late high-left (trials 11–20)	197	[191, 204]	185	[175, 194]
Early high-right (trials 1–10)	193	[187, 200]	179	[170, 188]
Late high-right (trials 11–20)	198	[190, 206]	185	[172, 197]

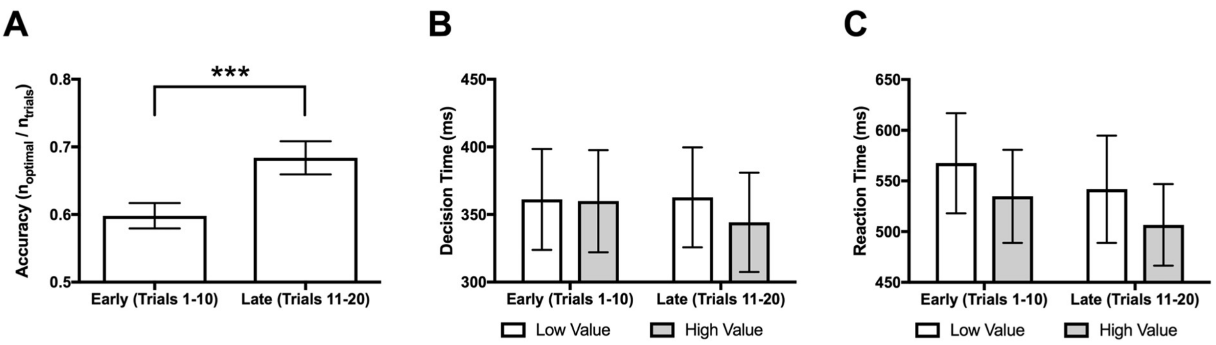

Fig. 2. Behavioral means and 95% confidence intervals. (a) Accuracy, defined as the proportion of trials the optimal choice was made, improved over time. ***$p < 0.001$. (b) Decision time to higher- and lower-valued choice stimuli, both early and late in a block. (c) Reaction time to probes appearing at locations of higher-valued stimuli and lower-valued stimuli both early and late in a block. Although responses to probes at locations of higher-valued stimuli were faster, the advantage did not change over time.

260–340 ms following feedback onset at channel FCz. Specifically, the grand average difference wave (all conditions) reached half of its maximum value at both 260 ms and 340 ms post feedback.

Thus, our analysis resulted in two N2pc peak amplitudes and two reward positivity peak amplitudes for each participant (one for each trial grouping within a block: early/late). The existence of the N2pc and reward positivity at each of these trial groupings was tested using single-samples *t*-tests of these peaks (as defined above) against zero (Krigolson and Holroyd, 2007; Rodríguez-Fornells et al., 2002). For both ERPs (N2pc and reward positivity) peaks at different trial groupings (early/late) were compared using paired-samples *t*-tests. All error bars and error measures reflect 95% confidence intervals (Loftus and Masson, 1994; Masson and Loftus, 2003). For all *t*-tests we computed Cohen's *d* as follows:

$$d = \frac{M_{\text{diff}}}{s_{\text{diff}}}$$

where M_{diff} and s_{diff} are the mean and standard deviation of the difference scores (or peaks in the case of the ERP existence tests) respectively (see Cumming, 2014).

3. Results

3.1. Behavioral results

3.1.1. Accuracy

We defined accuracy as the proportion of trials that a participant selected the higher-valued option (i.e., made the best choice). Mean accuracy improved from early in a block ($M = 0.61$, 95% CI [0.59, 0.62]) to late in a block ($M = 0.69$, 95% CI [0.68, 0.71]): $t(15) = 9.70$, $p < 0.001$, Cohen's $d = 2.71$. See Fig. 2a.

3.1.2. Decision

Decision times were faster in the second half of a block, $F(1,15) = 5.61$, $p = 0.03$, $\eta_p^2 = 0.22$. Decision times were also faster when the higher-valued stimulus was chosen, $F(1,15) = 9.99$, $p = 0.0065$, $\eta_p^2 = 0.35$. Finally, there was an interaction between trial (early/late) and chosen stimulus (high/low), $F(1,15) = 6.16$, $p = 0.025$ $\eta_p^2 = 0.29$ – value had more of an impact later in a block. See Fig. 2b and Table 2.

3.1.3. Probes

Reaction times were faster for probes at locations of higher-valued stimuli than for probes at locations of lower-valued stimuli (main effect of value), $F(1,15) = 23.26$, $p < 0.001$, $\eta_p^2 = 0.65$ (Fig. 2c, Table 2). However, this reaction time benefit did not change with learning – there was no interaction between trial (early/late) and value (higher/lower), $F(1,15) = 0.04$, $p = 0.84$, $\eta_p^2 = 0.003$. Interestingly, there was a main effect of trial (early/late), $F(1,15) = 17.19$, $p < 0.001$, $\eta_p^2 = 0.54$: reaction times later in a block were faster than reaction times early in a block, regardless of probe location. These results did not change if we used only the first and last catch trial within each block.

Table 2
Decision times and probe reaction times for higher- and lower-valued stimuli, early and late in a block.

	Lower-valued stimuli		Higher-valued stimuli	
	RT (ms)	95% CI	RT (ms)	95% CI
Early *decision* (trials 1–10)	361	[324, 399]	360	[322, 398]
Late *decision* (trials 11–20)	363	[326, 400]	344	[307, 381]
Early *probe* (trials 1–10)	567	[518, 617]	534	[488, 580]
Late *probe* (trials 11–20)	542	[489, 595]	507	[466, 547]

3.2. Electroencephalographic results

3.2.1. ERP to bilateral visual stimuli: The N2pc

An analysis of the difference wave (contralateral minus ipsilateral) locked to the onset of the bilateral visual stimuli (the two choices) revealed an ERP component in the N2pc time range (220–300 ms), both early and late in a block, $t(15) = 2.73$, $p = 0.016$, Cohen's $d = 0.68$ (early), and $t(15) = 4.24$, $p < 0.01$, Cohen's $d = 1.06$ (late) – see Figs. 3 and A1b. Furthermore, the amplitude of the N2pc increased from early to late in a block (early: −0.24 μV, 95% CI [−0.42, −0.05], late: −0.52 μV, 96% CI [−0.78, −0.26]), $t(15) = 2.69$, $p = 0.017$, Cohen's $d = 0.67$. The N2pc early in a block was dependent on behavior: trials in which the higher-valued stimulus was chosen elicited the usual negative response (−1.12 μV, 95% CI [−1.63, −0.62], $t(15) = 4.72$, $p < 0.001$, Cohen's $d = 1.18$). However, trials in which the lower-valued stimulus was chosen elicited the reverse response – that is, an N2pc was observed contralateral to lower-valued stimuli (0.68 μV, 95% CI [0.35, 1.01], $t(15) = 4.38$. $p < 0.001$, Cohen's $d = 1.09$). See Fig. 5.

3.2.2. ERP to feedback: The reward positivity

An analysis of difference waves (wins minus losses) locked to the onset of feedback revealed an ERP component with a latency (280–330 ms) and scalp distribution (maximal at FCz) consistent with a reward positivity, both early and late in a block, $t(15) = 5.82$, $p < 0.01$, Cohen's $d = 1.45$ (early), and $t(15) = 4.22$, $p < 0.01$, Cohen's $d = 1.06$ (late) – see Figs. 4 and A1a (also see Holroyd and Coles, 2002; Miltner et al., 1997). Furthermore, the reward positivity was reduced late in a block, after learning had occurred (early: 4.68 μV, 95% CI [2.97 6.40], late: 3.10 μV, 95% CI [1.54, 4.67]), $t(15) = 2.62$, $p = 0.019$, Cohen's $d = -0.65$.

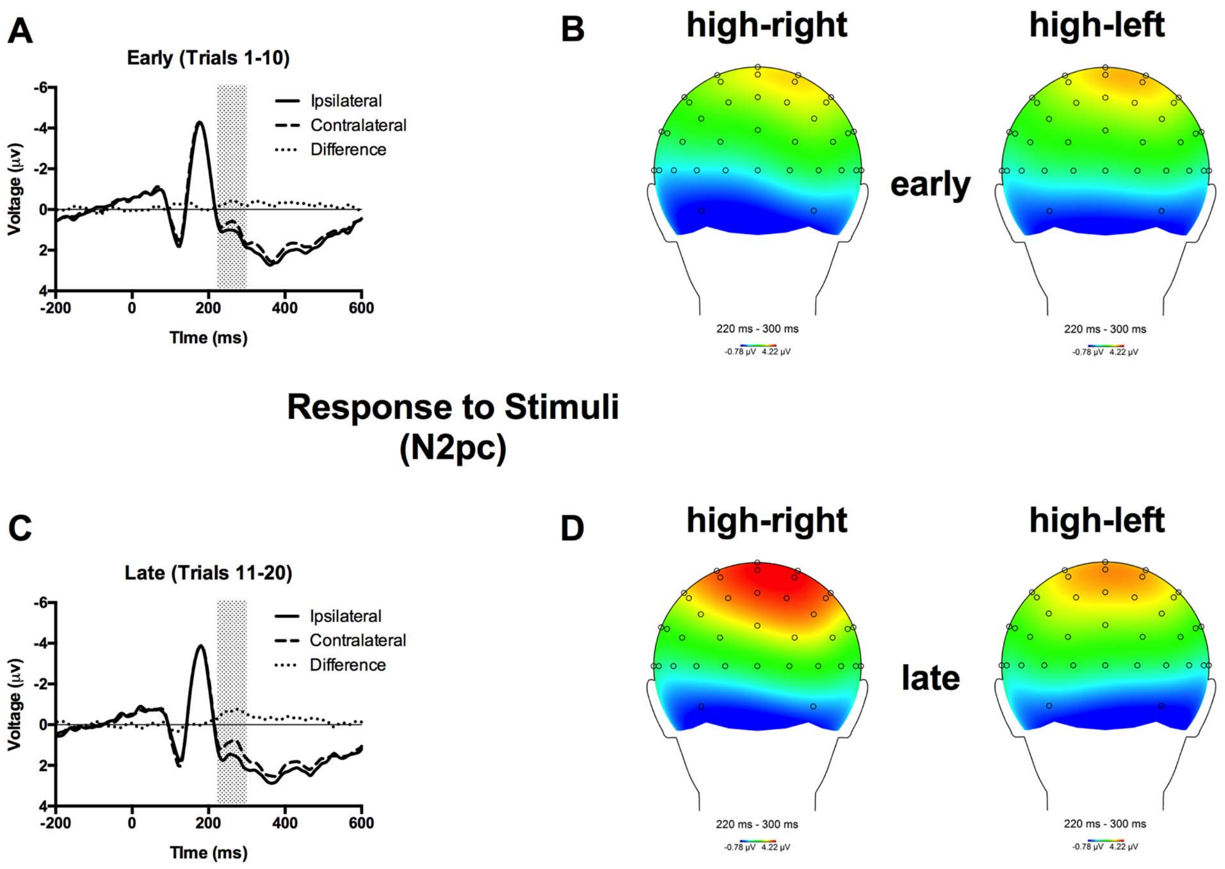

Fig. 3. Event-related brain potential (ERP) responses to choice stimuli and associated conditional scalp topographies both early (a, b) and late (c, d) in learning. The waveforms (a, c) reflect the average of channels PO7 and PO8, depending on whether the channel was ipsilateral or contralateral to the higher-valued option (see topographies). Negative is plotted up by convention. Shaded regions show the window of analysis for the N2pc. Note that although the N2pc difference waves were maximal at PO7/8, the conditional topographies (b, d) were maximal at O1 and O2.

4. Discussion

The goal of the current experiment was to provide evidence that humans engage an RL system when faced with multidimensional choices, and that learning in such environments is guided by attentional biases toward relevant dimensions (the sparsity bias). On each trial, we presented participants with two complex stimuli, one of which was more likely to yield a win when chosen because of a feature it possessed within a single dimension, i.e., the relevant dimension. In line with previous work (Gershman et al., 2010; Leong et al., 2017; Niv et al., 2015; Wilson and Niv, 2012) our participants learned this task at a rate comparable to a hybrid model – a model combining simple RL rules with an attentional bias toward those dimensions believed to be predictive of the outcome. In addition to this performance result, our attentional probe result supported the hypothesis that humans employ the sparsity bias when learning to choose from among multidimensional choice stimuli. Specifically, participants responded faster to unexpected attentional probes that appeared within higher-valued choices compared to lower-valued choices. Since only a single dimension of our choice stimuli was ever predictive of wins and losses, this suggests that selective attention was engaged within relevant dimensions more so than within irrelevant dimensions (recall that selective attention is the proposed mechanism behind the sparsity bias).

Our ERP results were generally consistent with our behavioral results: the N2pc, an ERP marker of selective attention, was enhanced for higher-valued stimuli relative to lower-valued stimuli. We used the N2pc to track the deployment of attention in the bilateral visual displays. Early on, before participants had learned which stimulus was higher in value, the N2pc appeared contralateral to whichever stimulus was eventually chosen, regardless of correctness (Fig. 5). Later on in a block, however, the N2pc appeared contralateral to the optimal choice, indicating that participants deployed attention to the more valuable stimulus. Since only a single dimension was ever relevant, this supports (and is the first ERP evidence for) the existence of the sparsity bias. Our other ERP evidence concerned feedback processing, and focused on the reward positivity, an ERP component thought to index a reward prediction error (Holroyd and Coles, 2002). Not only did we observe a reward positivity both early and late in learning (suggesting the involvement of an RL system), we observed a learning-dependent decrease in the magnitude of the reward positivity. This decrease in the reward positivity mirrored the predicted reduction in reward prediction errors generated by a hybrid RL model, and is consistent with previous studies on learning-related changes in the reward positivity (e.g., Krigolson et al., 2009, 2014).

Interestingly, and somewhat in contrast to our N2pc result, we did not observe a change in the value-based reaction time benefit across time: participants were faster to respond to probes at higher-valued locations both early and late in a block. Conversely, the N2pc effect (which was also present both early and late in a block) increased with learning. This seeming conflict between our ERP and behavioral data is curious, especially if the same mechanism (selective attention) underlies both effects. One possible explanation is that in addition to selective attention, the N2pc may be sensitive to other factors, such as the number of distractors (e.g., Luck et al., 1997; Luck and Hillyard, 1994b). Here, for example, there may be two types of distractors requiring suppression: the lower-valued features of the relevant dimension, and the non-relevant dimensions. It is also worth mentioning that although the probe reaction time benefit did not change over time (Fig. 2c), the effect of value on actual decision times did (Fig. 2b), more closely mirroring our N2pc results.

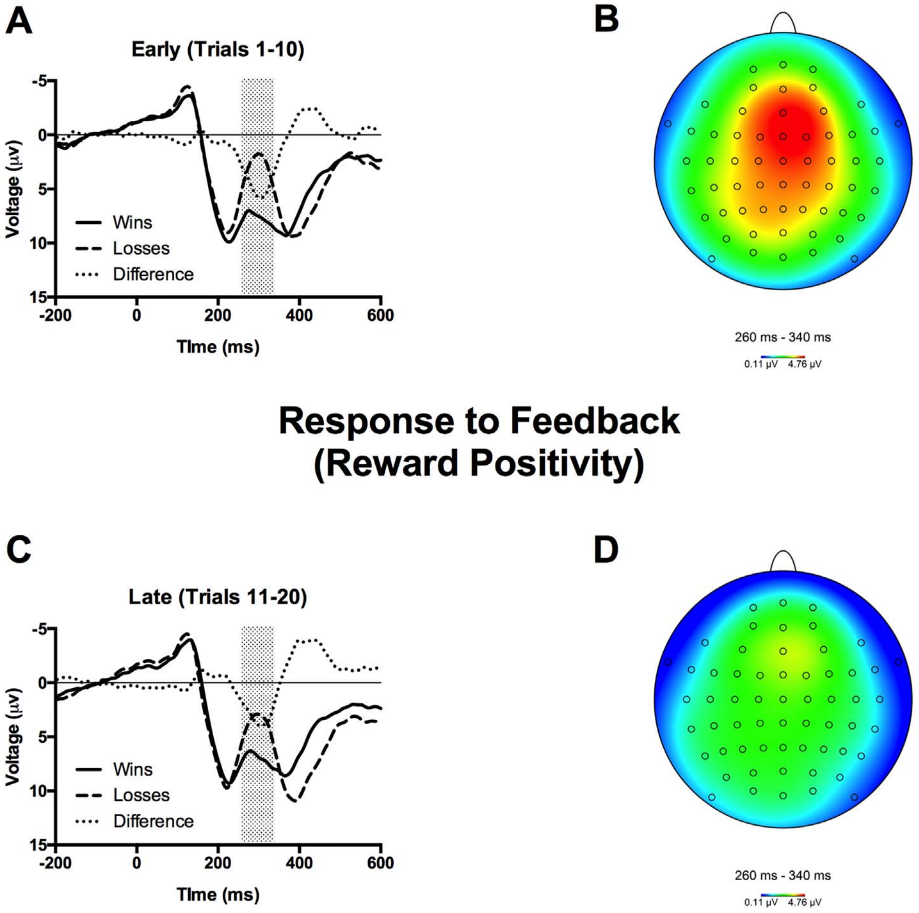

Fig. 4. Event-related brain potential (ERP) responses and difference wave scalp topographies both early (a, b) and late (c, d) in learning. Negative is plotted up by convention. Analysis was done at channel FCz. Shaded areas indicate where the reward positivities were computed.

In general, the behavioral and ERP results presented here are in line with a growing body of literature suggesting that our decisions are based on an attentionally-filtered version of the world. While selective attention is traditionally seen as the answer to our limited processing abilities (e.g., Kahneman, 1973), it may actually provide computational benefits beyond – or instead of – dealing with constrained resources (Dayan et al., 2000). In other words, even if we were to possess infinite processing resources, there might still be a performance advantage to narrowing the basis of our decision making to only those dimensions believed to be most relevant. Thus, learning and maintaining an accurate Bayesian model of the world is beneficial, and it has been suggested that we possess specialized systems for doing so (Yu and Dayan, 2005). Furthermore, such a Bayesian model may account for (and indeed, underlie) our attentional control system (Yu et al., 2009).

It is worth noting that our results, while consistent with the sparsity bias, are open to alternative interpretations. For example, although we have demonstrated attentional shifts related to relevant dimensions and valuable features, these attentional biases may not have been driving learning. In other words, participants may have learned the appropriate responses in our task independent of attentional processes, and may only have incidentally directed their attention to the higher-valued choice. However, we feel it more likely that participants were guiding their learning through sparsity since, in line with previous work (Gershman et al., 2010; Leong et al., 2017; Niv et al., 2015; Wilson and

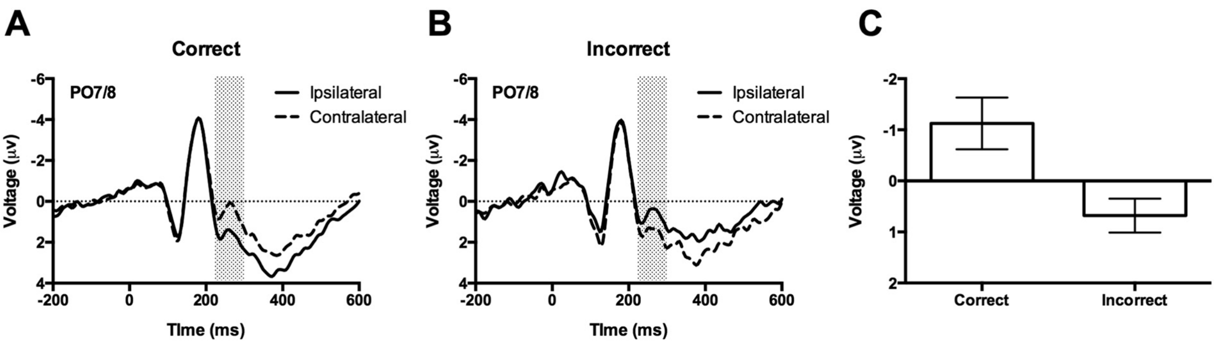

Fig. 5. Event-related brain potential (ERP) responses to choice stimuli prior to correct (a) and incorrect (b) choices. (c) Mean and 95% confidence intervals for the N2pc amplitude in response to choice stimuli prior to correct and incorrect choices. Note that only trials early in a block are shown and that correct trials are defined as trials in which the optimal choice was made (i.e., the higher-valued stimulus was selected), regardless of the actual outcome. Trials in which the lower-valued stimulus was selected were characterized by a "reverse" N2pc, suggesting that participants selectively attended to the lower-valued stimulus prior to making an incorrect response.

Niv, 2012), our participants were informed about the sparse nature of the environment. In addition to the above-mentioned limitation in our interpretation, it should be mentioned that participants in the current study were not playing for money, only points. Both monetary and point rewards have been shown to elicit a reward positivity, although a case can be made for paying participants (e.g., it results in a higher-amplitude signal; see Van den Berg et al., 2012; though also see Ma et al., 2014). Finally, although we are confident that we collected a sufficient number of trials for each condition (Table 1), our sample size was relatively small by current ERP research standards.

Both the behavioral and neuroimaging data presented here suggest that the human RL system can deal with multidimensional choices when guided by the sparsity bias: learning is focused on those dimensions believed to be most predictive of reward. The mechanism by which sparsity is implemented appears to be an attentional bias, as indexed by the N2pc component of the human brain ERP. Thus, a seemingly simple approach like RL may be capable of learning complex tasks if it operates on an attentionally-filtered version of the world.

Acknowledgements

This research was supported by Natural Sciences and Engineering Research Council of Canada graduate funding (first author) and Discovery grant (last author: RGPIN 2016-0943).

Appendix A

A.1. Computational models

Our computational models were closely based on those of Gershman et al. (2010; also see Leong et al., 2017; Niv et al., 2015; Wilson and Niv, 2012). What follows is a summary of the different models along with some implementation details. See Figs. A1 and A2 for model outputs.

A.1.1. RL model

The RL model maintained values associated with each possible combination of properties (a total of $3^3 = 27$ combinations/values) that were used to determine which combination was chosen according to the softmax algorithm (see Action Selection, below). The value associated with a chosen combination was updated following feedback by computing a prediction error δ – the difference between the reward that was received and the value v for that combination. Thus, if choosing combination i resulted in feedback R then a prediction error was computed according to

$$\delta = R - v_i$$

where $R = 1$ for a win and $R = -1$ for a loss. The prediction error was then used to update the value as follows:

$$v_i = v_i + \alpha\delta$$

Prediction errors were scaled by the learning rate ($\alpha = 0.8$, which maximized performance for this model). While this model learned optimal responses, it did so slowly because it could only deal with specific combinations of features and could not generalize between combinations sharing a common feature.

A.1.2. Bayesian model

Our Bayesian model used Bayes' rule to compute the expected value of each choice, V_1 and V_2. Similar to the RL model, the Bayesian model used the softmax algorithm (on expected value) for action selection. Computing expected values required maintaining, for each possible feature within a dimension, a prior (the likelihood that a feature is the target or higher-valued feature of the two being presented) and joint probability (the likelihood that a win will result if the feature is selected). In particular, the prior probability that feature f is the target was computed by

$$p(\textit{target is } f \mid \textit{ history}) \propto \prod_t p(R = r_t \mid \textit{history}) \textit{ if } f \textit{ were the target}$$

That is, the likelihood that f is the target (higher-valued feature) in a trial given the history is the likelihood of getting the particular configuration of wins and losses received for the history of states and choices made if the target were indeed f. We can compute this probability for each trial: if f is selected in a trial, the probability of a win would be 0.725 (average of the higher-valued target probabilities 0.95 and 0.5), and if not selected, it would be 0.275 (average of 0.05 and 0.5), as defined by the task structure.

The likelihood that choosing f would lead to a win can be reduced to

$$p(\textit{win} \mid \textit{selecting } f) = \frac{\textit{number of times selecting } f \textit{ led to a win}}{\textit{number of times } f \textit{ was selected}}$$

Then, the expected values associated with the displayed features were calculated as:

$V_f = p(\textit{win}|f)p(\textit{target is } f|\textit{ history}) - p(\textit{loss}|f)p(\textit{target is not } f|\textit{ history})$ where the probability of a loss when selecting feature f is computed analogously to the probability of a win when selecting feature f as described above. Finally, the expected value of each choice was computed as the sum of the expected values of the features making up that choice. For example, for choice i:

$$V_i = \sum_{f=1}^{3} V_f$$

A.1.3. Hybrid model

To model our ERP and behavioral results we used a hybrid RL model, as described by Gershman et al. (2010; also see: Leong et al., 2017; Niv et al., 2015; Wilson and Niv, 2012). The hybrid RL model was able to generalize across different complex stimuli by maintaining a value associated with each dimension's possible features. The total value associated with each choice could then be computed as a weighted sum of feature values. As with the RL model, a prediction error was computed following feedback (the difference between expected reward and actual reward) and used to update stored values (for features, not for combinations of features as in the RL model). For ease of comparison, the same learning rate ($\alpha = 0.8$) was used as with the RL model. This learning rate also resulted in maximal performance from the Hybrid model (and in fact, performance for both models was maximal for learning rates ranging from 0.7 to 0.9). Importantly, the hybrid model biased its learning according to what it believed to be the relevant dimension for the current block. Similar to the Bayesian model, it maintained beliefs over dimensions (shape, color, and texture) that were updated following feedback. Thus, for each dimension the model maintained a belief ϕ for each dimension d as per:

$$\phi_d = \frac{P(d \mid history)}{\sum_d P(d \mid history)}$$

where posteriors over each dimension were the marginal probabilities computed across each dimension's possible features. The dimensional beliefs, ϕ, were then interpreted as attentional weights, and used to bias the reinforcement learning rule where the value of each chosen feature f (belonging to dimension d) was updated according to:

$$v_i(f) = v_i(f) + \alpha\delta\phi_d$$

where prediction errors δ were computed as before. For example, if choosing a red triangle with dots was rewarded, then the values associated with red, triangle, and dots were updated and biased by the beliefs that either color, shape, or texture was the relevant dimension. Note that here we updated a value associated with each feature (as opposed to each combination of features, as in the RL model). See Fig. A1c for how mean prediction errors δ decreased over the course of a block. To show changes in the attentional weights over time, we plotted the mean maximum (strongest) dimensional weight ϕ (Fig. A1d) on each trial.

A.1.4. Action selection

Consistent with human decision making (Daw et al., 2006) our models used the softmax algorithm to select actions (Sutton and Barto, 1998; Gershman et al., 2010). This algorithm took the values associated with different actions and computed the likelihoods of selecting those actions. A temperature parameter τ allowed some control over the stochasticity of the model's choices. We set $\tau = 0.2$ for the RL and hybrid models, and $\tau = 0.001$ (low exploration) for the Bayesian model. These temperatures maximized performance for each model.

$$P(action\ i\ is\ selected) = \frac{e^{v_i/\tau}}{e^{v1/\tau} + e^{v2/\tau}}$$

Simulations of the above three models are shown in Fig. A2, where the probability of making the correct (higher-probability of winning) choice across 40 blocks is shown over the 20 trial block length where, in each block, the target dimension and feature property outcome probabilities remained constant. Similar to previous work (Gershman et al., 2010), the Bayesian model performed best, followed by the hybrid model, and lastly by the reinforcement learning model. The Bayesian model is the statistically optimal model for the task and so must make the best decisions by definition. The hybrid model, a dimensionally-attentive version of the RL model, makes better responses than the RL algorithm alone.

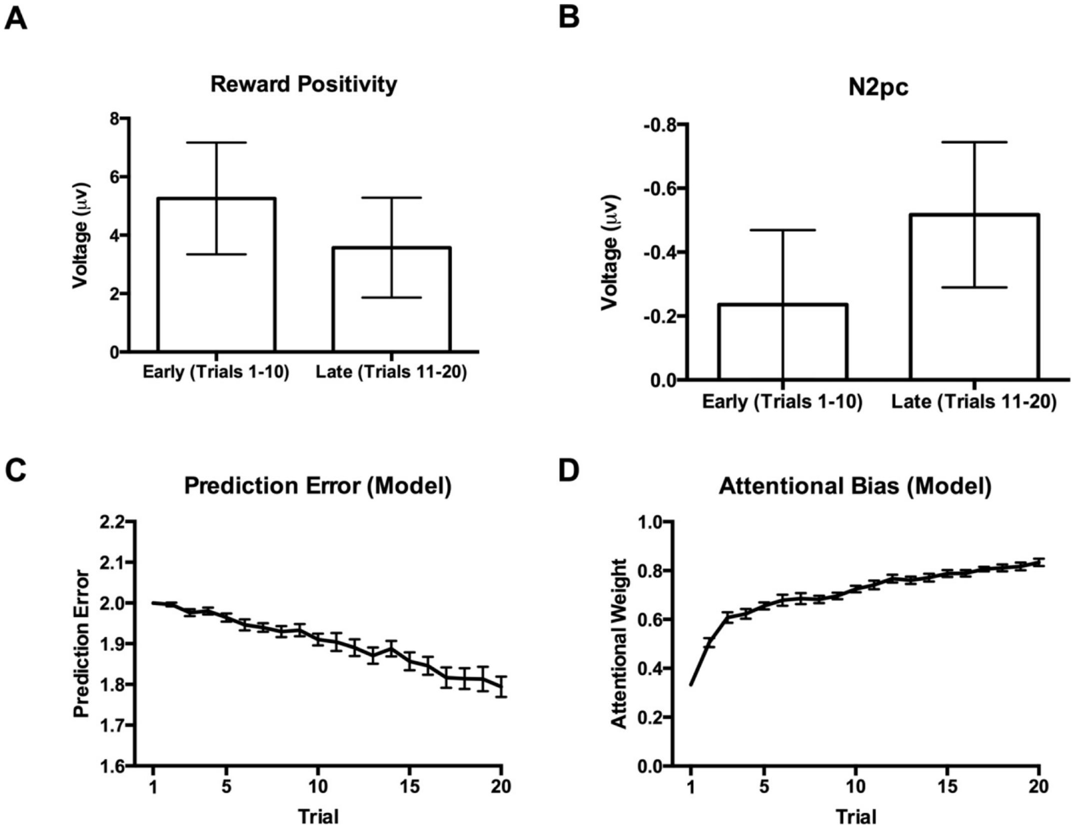

Fig. A1. Mean and 95% confidence intervals for the human event-related brain potential (ERP) peaks and mean and 95% confidence intervals for the model predictions. (a) The reward positivity was present both early and late in learning, and diminished over time. Positive is plotted up to emphasize that this is a positive ERP component. (b) The N2pc was also present both early and late in learning, and increased with learning. Negative is plotted up to emphasize that this is a negative ERP component. These ERP results mirrored the prediction errors (c) and attentional biases (d) generated by a hybrid model (RL with Bayesian weightings).

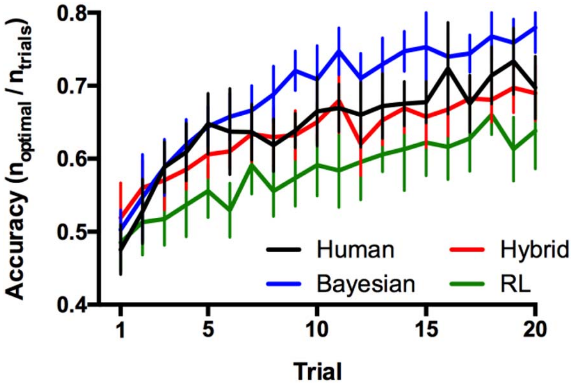

Fig. A2. Mean model and human performance, with 95% confidence intervals. Humans performed better than a simple reinforcement learning model (RL), but not optimally (Bayesian). A hybrid approach (Hybrid) in which learning was focused on attended dimensions provided more overlap to our human data.

References

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.

Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10 (4), 433–436. http://dx.doi.org/10.1163/156856897X00357.

Chase, H.W., Kumar, P., Eickhoff, S.B., Dombrovski, A.Y., 2015. Reinforcement learning models and their neural correlates: an activation likelihood estimation meta-analysis. Cogn. Affect. Behav. Neurosci. 15 (2), 435–459.

Cumming, G., 2014. The new statistics: why and how. Psychol. Sci. 25 (1), 7–29. http://dx.doi.org/10.1177/0956797613504966.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441 (7095), 876–879. http://dx.doi.org/10.1038/nature04766.

Dayan, P., Kakade, S., Montague, P.R., 2000. Learning and selective attention. Nat. Neurosci. 3, 1218–1223.

Eimer, M., 1996. The N2pc component as an indicator of attentional selectivity. Electroencephalogr. Clin. Neurophysiol. 99 (3), 225–234.

Eimer, M., Kiss, M., 2008. Involuntary attentional capture is determined by task set: evidence from event-related brain potentials. J. Cogn. Neurosci. 20 (8), 1423–1433. http://dx.doi.org/10.1162/jocn.2008.20099.

Gershman, S.J., Cohen, J.D., Niv, Y., 2010. Learning to selectively attend. In: Ohlsson, S., Catrambone, R. (Eds.), Proceedings of the 32nd Annual Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 1270–1275.

Hickey, C., Di Lollo, V., McDonald, J.J., 2009. Electrophysiological indices of target and distractor processing in visual search. J. Cogn. Neurosci. 21 (4), 760–775. http://dx.doi.org/10.1162/jocn.2009.21039.

Hillyard, S.A., Anllo-Vento, L., 1998. Event-related brain potentials in the study of visual selective attention. Proc. Natl. Acad. Sci. 95 (3), 781–787.

Holroyd, C.B., Coles, M.G., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109 (4), 679.

Holroyd, C.B., Krigolson, O.E., 2007. Reward prediction error signals associated with a modified time estimation task. Psychophysiology 44 (6), 913–917. http://dx.doi.org/10.1111/j.1469-8986.2007.00561.x.

Holroyd, C.B., Nieuwenhuis, S., Yeung, N., Cohen, J.D., 2003. Errors in reward prediction are reflected in the event-related brain potential. Neuroreport 14 (18), 2481.

Holroyd, C.B., Pakzad-Vaezi, K.L., Krigolson, O.E., 2008. The feedback correct-related positivity: sensitivity of the event-related brain potential to unexpected positive feedback. Psychophysiology 45 (5), 688–697. http://dx.doi.org/10.1111/j.1469-8986.2008.00668.x.

Holroyd, C.B., Krigolson, O.E., Lee, S., 2011. Reward positivity elicited by predictive cues. Neuroreport 22 (5), 249.

Kahneman, D., 1973. Attention and Effort. Prentice-Hall Inc.

Kiss, M., Van Velzen, J., Eimer, M., 2008. The N2pc component and its links to attention shifts and spatially selective visual processing. Psychophysiology 45 (2), 240–249. http://dx.doi.org/10.1111/j.1469-8986.2007.00611.x.

Krigolson, O.E., Holroyd, C.B., 2007. Predictive information and error processing: the role of medial-frontal cortex during motor control. Psychophysiology 44 (4), 586–595. http://dx.doi.org/10.1111/j.1469-8986.2007.00523.x.

Krigolson, O.E., Pierce, L.J., Holroyd, C.B., Tanaka, J.W., 2009. Learning to become an expert: reinforcement learning and the acquisition of perceptual expertise. J. Cogn. Neurosci. 21 (9), 1833–1840.

Krigolson, O.E., Hassall, C.D., Handy, T.C., 2014. How we learn to make decisions: rapid propagation of reinforcement learning prediction errors in humans. J. Cogn. Neurosci. 26 (3), 635–644. http://dx.doi.org/10.1162/jocn_a_00509.

Kruschke, J.K., Kappenman, E.S., Hetrick, W.P., 2005. Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. J. Exp. Psychol. Learn. Mem. Cogn. 31 (5), 830–845. http://dx.doi.org/10.1037/0278-7393.31.5.830.

Leong, Y.C., Radulescu, A., Daniel, R., DeWoskin, V., Niv, Y., 2017. Dynamic interaction between reinforcement learning and attention in multidimensional environments. Neuron 93 (2), 451–463. http://dx.doi.org/10.1016/j.neuron.2016.12.040.

Loftus, G.R., Masson, M.E.J., 1994. Using confidence intervals in within-subject designs. Psychon. Bull. Rev. 1 (4), 476–490.

Luck, S.J., Hillyard, S.A., 1994a. Electrophysiological correlates of feature analysis during visual search. Psychophysiology 31 (3), 291–308.

Luck, S.J., Hillyard, S.A., 1994b. Spatial filtering during visual search: evidence from human electrophysiology. J. Exp. Psychol. Hum. Percept. Perform. 20 (5), 1000–1014.

Luck, S.J., Fan, S., Hillyard, S.A., 1993. Attention-related modulation of sensory-evoked brain activity in a visual search task. J. Cogn. Neurosci. 5 (2), 188–195. http://dx.doi.org/10.1162/jocn.1993.5.2.188.

Luck, S.J., Girelli, M., McDermott, M.T., Ford, M.A., 1997. Bridging the gap between monkey neurophysiology and human perception: an ambiguity resolution theory of visual selective attention. Cogn. Psychol. 33 (1), 64–87. http://dx.doi.org/10.1006/cogp.1997.0660.

Ma, Q., Jin, J., Meng, L., Shen, Q., 2014. The dark side of monetary incentive: how does extrinsic reward crowd out intrinsic motivation. Neuroreport 25 (3), 194. http://dx.doi.org/10.1097/WNR.0000000000000113.

Mackintosh, N.J., 1975. A theory of attention: variations in the associability of stimuli with reinforcement. Psychol. Rev. 82 (4), 276.

MacLeod, C., Mathews, A., Tata, P., 1986. Attentional bias in emotional disorders. J. Abnorm. Psychol. 95 (1), 15.

Masson, M.E., Loftus, G.R., 2003. Using confidence intervals for graphically based data interpretation. Can. J. Exp. Psychol. 57 (3), 203.

McDonald, J.J., Green, J.J., Jannati, A., Di Lollo, V., 2013. On the electrophysiological evidence for the capture of visual attention. J. Exp. Psychol. Hum. Percept. Perform. 39 (3), 849–860. http://dx.doi.org/10.1037/a0030510.

Mill, J.S., 1863. Utilitarianism. Parker, Son, and Bourn, London, U. K.

Miltner, W.H.R., Braun, C.H., Coles, M.G.H., 1997. Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a “generic” neural system for error detection. J. Cogn. Neurosci. 9 (6), 788–798. http://dx.doi.org/10.1162/jocn.1997.9.6.788.

Müller, H.J., Reimann, B., Krummenacher, J., 2003. Visual search for singleton feature targets across dimensions: stimulus- and expectancy-driven effects in dimensional weighting. J. Exp. Psychol. Hum. Percept. Perform. 29 (5), 1021–1035. http://dx.doi.org/10.1037/0096-1523.29.5.1021.

Niv, Y., 2009. Reinforcement learning in the brain. J. Math. Psychol. 53 (3), 139–154. http://dx.doi.org/10.1016/j.jmp.2008.12.005.

Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., Wilson, R.C., 2015. Reinforcement learning in multidimensional environments relies on attention mechanisms. J. Neurosci. 35 (21), 8145–8157. http://dx.doi.org/10.1523/JNEUROSCI.2978-14.2015.

Oliveira, F.T.P., McDonald, J.J., Goodman, D., 2007. Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. J. Cogn. Neurosci. 19 (12), 1994–2004. http://dx.doi.org/10.1162/jocn.2007.19.12.1994.

Pearce, J.M., Hall, G., 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. 87 (6), 532.

Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10 (4), 437–442. http://dx.doi.org/10.1163/156856897X00366.

Proudfit, G.H., 2015. The reward positivity: from basic research on reward to a biomarker for depression: the reward positivity. Psychophysiology. http://dx.doi.org/10.1111/psyp.12370.

Rodríguez-Fornells, A., Kurzbuch, A.R., Münte, T.F., 2002. Time course of error detection and correction in humans: neurophysiological evidence. J. Neurosci. 22 (22), 9990–9996.

Roesch, M.R., Esber, G.R., Li, J., Daw, N.D., Schoenbaum, G., 2012. Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. Eur. J. Neurosci. 35 (7), 1190–1200. http://dx.doi.org/10.1111/j.1460-9568.2011.07986.x.

Sambrook, T.D., Goslin, J., 2015. A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. Psychol. Bull. 141 (1), 213–235. http://

dx.doi.org/10.1037/bul0000006.

Schultz, W., 2013. Updating dopamine reward signals. Curr. Opin. Neurobiol. 23 (2), 229–238. http://dx.doi.org/10.1016/j.conb.2012.11.012.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Van den Berg, I., Shaul, L., Van der Veen, F.M., Franken, I.H.A., 2012. The role of monetary incentives in feedback processing: why we should pay our participants. Neuroreport 23 (6), 347. http://dx.doi.org/10.1097/WNR.0b013e328351db2f.

Wills, A.J., Lavric, A., Croft, G.S., Hodgson, T.L., 2007. Predictive learning, prediction errors, and attention: evidence from event-related potentials and eye tracking. J. Cogn. Neurosci. 19 (5), 843–854. http://dx.doi.org/10.1162/jocn.2007.19.5.843.

Wilson, R.C., Niv, Y., 2012. Inferring relevance in a changing world. Front. Hum. Neurosci. 5. http://dx.doi.org/10.3389/fnhum.2011.00189.

Yu, A.J., Dayan, P., 2005. Uncertainty, neuromodulation, and attention. Neuron 46 (4), 681–692. http://dx.doi.org/10.1016/j.neuron.2005.04.026.

Yu, A.J., Dayan, P., Cohen, J.D., 2009. Dynamics of attentional selection under conflict: toward a rational Bayesian account. J. Exp. Psychol. Hum. Percept. Perform. 35 (3), 700–717. http://dx.doi.org/10.1037/a0013553.