Behavioral/Cognitive

Goal-Directed and Habit-Like Modulations of Stimulus Processing during Reinforcement Learning

David Luque, Tom Beesley, Richard W. Morris, Bradley N. Jack, Oren Griffiths, Thomas J. Whitford, and Mike E. Le Pelley
School of Psychology, UNSW Australia, Sydney, New South Wales 2052, Australia

Recent research has shown that perceptual processing of stimuli previously associated with high-value rewards is automatically prioritized even when rewards are no longer available. It has been hypothesized that such reward-related modulation of stimulus salience is conceptually similar to an "attentional habit." Recording event-related potentials in humans during a reinforcement learning task, we show strong evidence in favor of this hypothesis. Resistance to outcome devaluation (the defining feature of a habit) was shown by the stimulus-locked P1 component, reflecting activity in the extrastriate visual cortex. Analysis at longer latencies revealed a positive component (corresponding to the P3b, from 550–700 ms) sensitive to outcome devaluation. Therefore, distinct spatiotemporal patterns of brain activity were observed corresponding to habitual and goal-directed processes. These results demonstrate that reinforcement learning engages both attentional habits and goal-directed processes in parallel. Consequences for brain and computational models of reinforcement learning are discussed.

Key words: attention; event-related potentials; goal-directed; habit; learning; reward

Significance Statement

The human attentional network adapts to detect stimuli that predict important rewards. A recent hypothesis suggests that the visual cortex automatically prioritizes reward-related stimuli, driven by cached representations of reward value; that is, stimulus–response habits. Alternatively, the neural system may track the current value of the predicted outcome. Our results demonstrate for the first time that visual cortex activity is increased for reward-related stimuli even when the rewarding event is temporarily devalued. In contrast, longer-latency brain activity was specifically sensitive to transient changes in reward value. Therefore, we show that both habit-like attention and goal-directed processes occur in the same learning episode at different latencies. This result has important consequences for computational models of reinforcement learning.

Introduction

Animals and humans learn to make specific responses to acquire rewards and avoid punishments. Awareness of an action's consequences allows the performer to rapidly and flexibly adapt their behavior when the value of those consequences changes. Such goal-directed behavior is especially evident in new learning. With repeated experience of an invariant response–reward relationship, however, behavior may become more reflexive or habitual. The expression of behavioral habits relies on the cached value of S–R (stimulus–response) links so that, when the S is perceived, the R is automatically elicited. Given that these S–R links do not incorporate information about the current value of the outcome, the resulting habitual behavior will be unaffected by recent changes in outcome value. Therefore, behaviors that are shown to be insensitive to posttraining alterations in the outcome value are attributed to the operation of the habit system, whereas rapid behavioral adaption to new outcome values is an index of the operation of the goal-directed system (Balleine and O'Doherty, 2010).

Recent research suggests that it is not merely overt behavioral responses that are modified by the formation of S–R habits, but rather that habits can also shape how the perceptual system allocates attention to (and thus prioritizes processing of) stimuli (Anderson, 2016; Le Pelley et al., 2016). Specifically, if attentional selection of a stimulus consistently yields high reward, then that stimulus will become more likely to capture attention and this pri-

Received Oct. 16, 2016; revised Jan. 5, 2017; accepted Jan. 31, 2017.

Author contributions: D.L., T.B., R.W.M., O.G., T.J.W., and M.E.L.P. designed research; D.L. and B.N.J. performed research; D.L. analyzed data; D.L., T.B., R.W.M., B.N.J., O.G., T.J.W., and M.E.L.P. wrote the paper.

This work was supported by the Australian Research Council (Discovery Projects DP140104394 and DP160103063 to T.J.W., M.E.L.P., and D.L. and to T.B. and D.L., respectively, Discovery Early Career Research Award (DE150100667 to O.G., and Future Fellowship FT100100260 to M.E.L.P. T.J.W. is also supported by Career Development Fellowship APP1090507 from the National Health and Medical Research Council of Australia. We thank Miranda Chilver and Esmeralda Paric for assistance with the experiments.

The authors declare no competing financial interests.

Correspondence should be addressed to David Luque, School of Psychology, UNSW Australia, Sydney, New South Wales 2052, Australia. E-mail: d.luque@unsw.edu.au.

DOI:10.1523/JNEUROSCI.3205-16.2017

 0270-6474/17/373009-09$15.00/0

Table 1. Experiment design

S–R–O mapping			Outcome value		
S	→ R	→ O	No Dev	Dev O^{high}	Dev O^{low}
S^{high}	→ R1	→ O^{high}	+100	0	+100
	→ R2	→ Error!	—	—	—
S^{low}	→ R1	→ Error!	—	—	—
	→ R2	→ O^{low}	+1	+1	0

Columns on the left side of the table show the S–R–O mappings that participants experienced. S^{high} and S^{low} denote high-value and low-value stimuli. R1 and R2 denote Response 1 and Response 2, which could be left or right button presses depending on the counterbalancing condition. O^{high} and O^{low} denote high-value and low-value outcomes. Given an incorrect response, no outcome was earned and an error feedback (Error!) was displayed instead. Columns on the right side of the table show the different point values associated with the possible outcomes during blocks in which neither outcome was devalued (No Dev), blocks in which the high-value outcome was devalued (Dev O^{high}), and blocks in which the low-value outcome was devalued (Dev O^{low}). Incorrect responses never earned any points (denoted by —).

oritization will persist even if reward is now omitted and selection of the stimulus becomes contrary to ongoing task goals. Consistent with the notion that attention can be a habit just like other forms of behavior, it has been shown that dopamine activity in the dorsolateral striatum, which is crucial for the operation of the habitual system (Yin et al., 2006; Balleine and O'Doherty 2010), is increased when previously rewarded stimuli are presented as distractors in attentional search tasks (Anderson et al., 2016).

Electrophysiological studies suggest that reward learning produces relatively low-level changes in visual processing. The occipital P1 is an early, visual event-related potential (ERP) component (~100 ms after stimulus onset) that is produced by activity in the extrastriate visual cortex and provides an index of stimulus salience (Heinze et al., 1994; Hillyard et al., 1998; Di Russo et al., 2002). Notably, stimuli paired with high reward elicit a larger P1 than stimuli paired with low reward (Hickey et al., 2010; MacLean and Giesbrecht, 2015). This suggests an influence of reward on attention. One interpretation is that the visual cortex uses the information provided by the habit-learning system to increase its sensitivity to high-reward stimuli, effectively increasing the (automatic) prioritization of these stimuli (MacLean and Giesbrecht, 2015; Anderson, 2016; Pearson et al., 2016).

Although existing data are consistent with the possibility that the habit system and early visual processing interact, habit-like stimulus prioritization in the attentional network has not been tested directly. The current experiment investigates this issue by assessing the sensitivity to outcome devaluation of the reward-related, stimulus-locked P1 effect. As noted earlier, the defining feature of an S–R habit is that it persists despite transient changes in the incentive value of the outcome (O). In an S–R–O reward-learning task, two stimuli were associated with responses that produced either a high-value outcome (O^{high}) or low-value outcome (O^{low}) (100 or 1 points respectively, where points had monetary value). Following previous research, we expect greater P1 for stimuli associated with O^{high} (termed S^{high}) than O^{low} (S^{low}; Table 1). Critically, in some blocks, a specific outcome (either O^{high} or O^{low}) was devalued so that, during that block, it was worth zero points (Table 1). If the reward-related P1 effect reflects modulations by a habit-like signal, it should be unaffected by this devaluation. In contrast, nonhabitual (i.e., goal-directed) behavior will be sensitive to the value of the goal (outcome) by definition, so activity relating to goal-directed processes should differ between blocks with O^{high} devalued and blocks with O^{low} devalued.

Materials and Methods

Participants and apparatus. Twenty-four healthy adults (mean ± SD; age, 27 ± 8 years; two left-handed; 14 male) took part in exchange for a monetary reward, which was dependent on the number of points earned at the end of the experiment ($30 AUD ± $2 AUD). All participants signed an informed consent form approved by the Human Research Ethics Advisory Panel (Psychology) of UNSW Australia and were treated in accordance with the Helsinki declaration. All reported normal or corrected-to-normal vision.

Participants were tested individually in a cubicle using a standard PC with a 23-inch monitor (1920 × 1280 resolution, 60 Hz refresh rate) positioned ~100 cm from the participant. Stimulus presentation was controlled with MATLAB (The MathWorks) using the Psychophysics Toolbox extensions (Kleiner et al., 2007). Responses were made using an Empirisoft DirectIN millisecond accurate keyboard. The apparatus for EEG recording is described below.

Stimuli and task. Participants engaged in a trial-by-trial reinforcement learning task. To implement the outcome devaluation manipulation, we framed the learning task in a cover story. Participants played the role of space traders, the mission of which was to trade cookies for diamonds with two aliens. These aliens had two different types of diamonds and one of them (the high-value outcome, O^{high}, worth 100 points) was more valuable than the other (the low-value outcome, O^{low}, worth 1 point). Participants were instructed that they should earn as many points as possible and, at the end of the experiment they would receive $1 AUD for every 1000 points they had earned. To implement the outcome devaluation manipulation, participants were instructed that each type of diamond needed to be stored in a different spaceship, but sometimes one of these spaceships had to return to Earth. Participants could not store a specific type of diamond while its corresponding spaceship was away, so in these blocks, the new value for that type of diamond was zero regardless of its usual value. The status of the spaceships (present/absent) only changed at the beginning of each block and participants were instructed accordingly by a message on the screen. In no-devaluation (No Dev) blocks, both diamonds could be stored and thus had their usual value. In devaluation of the high-value outcome (Dev O^{high}) blocks, O^{high} could not be stored and thus was worthless (it gave 0 points), whereas O^{low} could be stored and thus still had value (albeit low value). In devaluation of the low-value outcome (Dev O^{low}) blocks, O^{low} could not be stored and was worthless, whereas O^{high} still had (high) value (Table 1).

Reinforcement learning trials consisted of S–R–O sequences. Background screen color was set to black throughout the task. Each trial began with a central red fixation point presented for a random duration of 200–300 ms. One of two distinct "cookie" stimuli (S^{high} or S^{low}) then appeared at the center of the screen. These stimuli were easily discriminable colored circles (3.3° diameter), each containing several smaller circles of a different color (Fig. 1*A*). After 800 ms, pictures of two aliens (~2.8 × 4°) appeared at either side of the screen (6.9° from the center on the horizontal meridian). These aliens marked the two response options (R1 and R2): participants chose the alien to which they wished to give the cookie by pressing "q" or "p" to choose the left or right alien, respectively. After a response, the screen blanked for 300 ms before outcome feedback was presented in the center of the screen for 800 ms. If the response was incorrect (Table 1), the message "Error!" appeared in red font. If the response was correct, but the response time was slower than 2 s, the message "Time out, please respond faster" appeared. If the response was correct and faster than 2 s, one of two diamonds (O^{high} or O^{low}) appeared (size 2.6 × 2.3°). One of these diamonds was colored yellow and the other was blue. If, on any trial, participants responded before the aliens appeared, the message "Too fast! No diamond for you!" appeared for 3 s. For each participant, the specific pictures used as S^{high} and S^{low} (cookies), R1 and R2 (aliens), and O^{high} and O^{low} (diamonds), were assigned following a Latin square counterbalancing design. The left/right position of the aliens representing R1 and R2 was determined at random for each participant at the beginning of the experiment.

The last screen on each trial varied between blocks. In No Dev blocks, this screen showed the value of the diamond just earned ("+100" for O^{high} and "+1" for O^{low}). During devaluation blocks (Dev O^{high} or Dev O^{low} blocks), this information was hidden for all trials (including consumption trials; see below). As a cover story, before devaluation blocks, participants were told that information about the value of the diamonds was unavailable during these blocks due to solar interference (the final screen of each trial showed "??" instead of value during devaluation

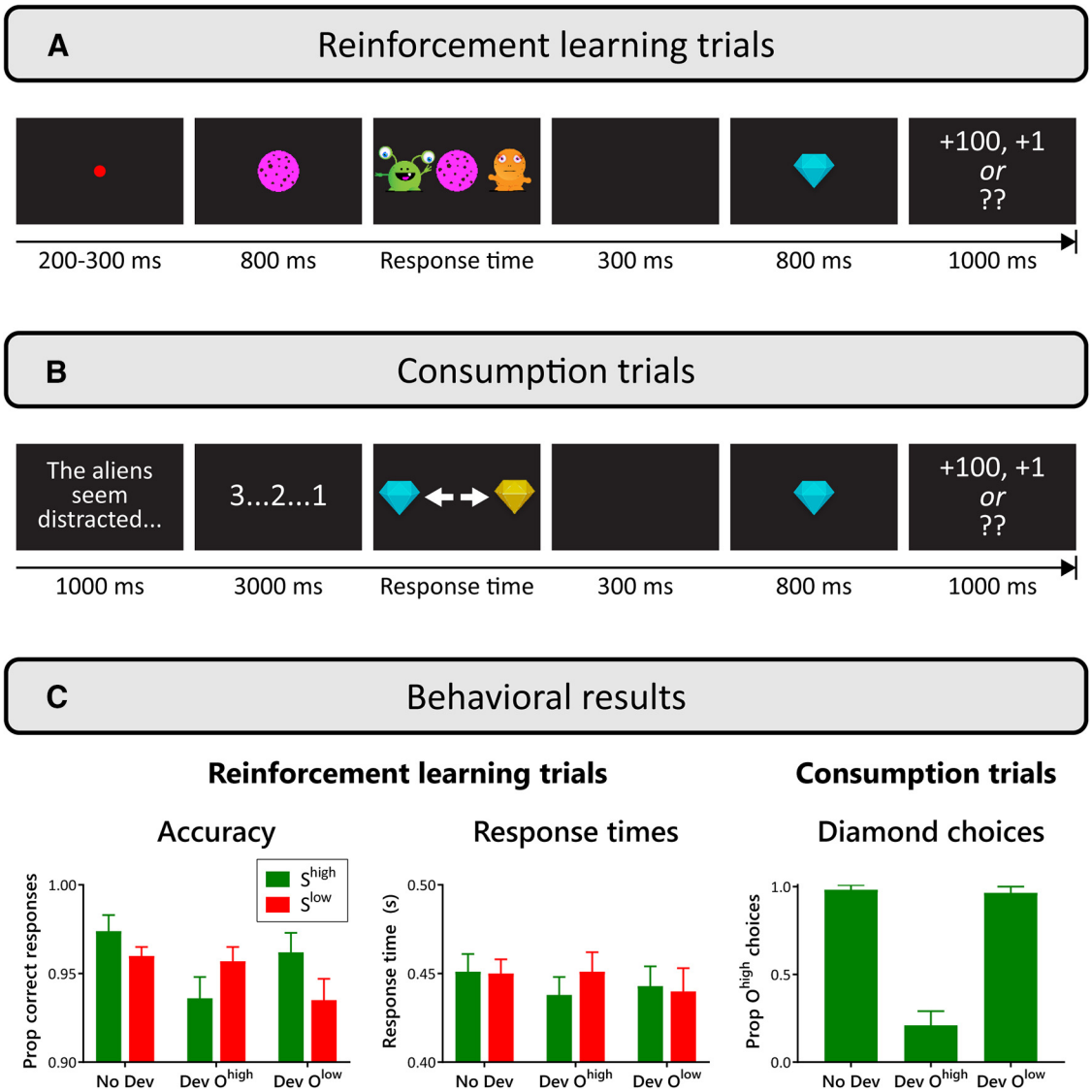

Figure 1. Paradigm and behavioral results. ***A***, Example of a learning trial. In this example, the participant earned the blue diamond by trading a pink cookie. The stimuli were counterbalanced between participants. The last screen for each trial showed the diamond value, although only in no-devalued blocks. During devalued blocks, two question marks were shown instead. Correct response to the stimulus (cookie) was rewarded with 100 or 1 points. ***B***, Example of a consumption trial. Participants chose between the two diamonds. Feedback and value screens were as in reinforcement learning trials. ***C***, Behavioral results showing response time and accuracy data for reinforcement learning trials and proportion of high-value outcome (diamond) choice for consumption trials. These results are divided into No Dev, Dev O^{high}, and Dev O^{low} blocks. Error bars indicate within-participant SEM.

blocks; Fig. 1*A*), but that this solar interference did not affect the value of the diamonds that they could earn. Trial-by-trial outcome values were hidden during devaluation blocks to avoid the updating of outcome values within the habit system (Tricomi et al., 2009; for a similar strategy, see de Wit et al., 2009; Gillan et al., 2014, 2015). The next trial began after an intertrial interval of 800 ms.

Consumption trials (Fig. 1*B*) were included in each block to provide a behavioral assay of participants' goal-directed behavior (Gillan et al., 2015). Instructions at the beginning of the experiment noted that sometimes the aliens become distracted and, on such occasions participants could take one of the two types of diamonds without trading it for cookies. On these consumption trials, the message "The aliens seem distracted . . ." appeared for 1000 ms, followed by a countdown (from 3 to 1) in the center of the screen (over 3 s). Then, the two diamonds were presented, one to the left and the other to the right (these positions were selected at random for each consumption trial), and participants selected the desired diamond using the "q" or "p" key. If participants responded before diamonds appeared, or response time was >2 s, "too fast" or "time out" messages appeared as in reinforcement learning trials. Outcome and value screens were also as for reinforcement learning trials.

Each block comprised 27 trials: 12 learning trials with S^{high}, 12 learning trials with S^{low}, and three consumption trials. Reinforcement learning trials within each block were presented in a random order. Consumption trials were presented as trial numbers 7, 14, and 21 within each block. The first block was always a No Dev block and, after that, the order for the different blocks followed a sequence: first a devaluation block in which the blue diamond (which was O^{high} for half of participants and O^{low} for the other half depending on counterbalance condition) was devalued, then a devaluation block in which the yellow diamond was devalued, and after that a No Dev block. This sequence was repeated 10 times, resulting in 31 blocks in total.

At the end of each block, the participants could see how many diamonds they had earned in the preceding block, the total points value of the diamonds earned in that block, and the total number of points they had earned in the experiment so far. After this screen, a new block was initiated with a screen informing the participant about what spaceship(s) was available for the next block.

EEG data acquisition. EEG was recorded with a Biosemi ActiveTwo system from 64 Ag/AgCl active electrodes placed according to the extended 10–20 system (FP1, FPz, FP2, AF7, AF3, AFz, AF4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, T7, C5, C3, C1, Cz, C2, C4, C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P9, P7, P5, P3, P1, Pz, P2, P4, P6, P8, P10, PO7, PO3, POz, PO4, PO8, O1, Oz, O2, Iz). Vertical eye movements were monitored by an electrode placed on the infraorbital ridge of the left eye; horizontal EOG was recorded by placing an electrode on the outer canthus of each eye. An electrode was also placed on the tip of the nose and on each mastoid. During data acquisition, the reference was composed of CMS and DRL sites and the sampling rate was 2048 Hz.

EEG preprocessing. Data preprocessing was performed using EEGLAB (Delorme and Makeig, 2004, RRID:SCR_007292). For data analyses, the EEG data were rereferenced offline to the average of both mastoids. For each participant, raw data were low-pass filtered using a phase-shift free Butterworth filter (12 dB/Oct slope) of 50 Hz and resampled at 250 Hz. An automatic bad channel detection algorithm was then applied using EEGLAB's pop_rejchan method (threshold = 5, method = kurtosis; Delorme and Makeig, 2004, RRID:SCR_007292). Independent components analysis (ICA) was applied to correct for ocular artifacts. After ICA, bad channels were interpolated (using the by-default spherical interpolation method) and stimulus-locked epochs extracted [1000 ms (−200 800), baselined (−100 0)]. To exclude any remaining artifacts, epochs with base-to-peak amplitude >100 mV were excluded from analyses (~6%). Finally, the first 4 blocks of the experiment [i.e., the initial No Dev block, and the following (first) sequence of the three critical block types: Dev O^{high}, Dev O^{low}, and No Dev] were not included in the ERP analysis to avoid any noise produced by lack of familiarity with the task because we were interested in reward-related effects at asymptotic levels of learning. Therefore, nine blocks of each type (No Dev, Dev O^{high}, and Dev O^{low}) were submitted for ERP analysis.

ERP analysis. Following MacLean and Giesbrecht (2015) and Hickey et al. (2010), we expected to find an effect of stimulus value (i.e., difference between activity elicited by S^{high} and S^{low}) in the occipital P1 component, peaking at some point within the time window 75–200 ms from stimulus onset. Because all stimuli were presented centrally, P1 activity was analyzed in the midline occipital electrode Oz (Smith et al., 2003). The P1 peak was defined as the largest positive peak between 75 and 200 ms after the onset of the stimulus at Oz (averaging across all conditions; MacLean and Giesbrecht, 2015). A time window of 70 ms around that peak was then selected for analysis. Because the P1 maximum amplitude was at 165 ms from stimulus onset, the P1 magnitude for each condition was defined as the mean EEG signal across the 130–200 ms time window. Repeated-measures ANOVA was used to assess the effect of stimulus value, outcome devaluation (i.e., differences between the No Dev, Dev O^{high}, and Dev O^{low} blocks) and their interaction on P1 magnitude.

We did not have specific, a priori predictions regarding ERPs during the remaining 600 ms in which the stimuli were presented (i.e., from 200 to 800 ms after stimulus onset). On the grounds that goal-directed brain processes probably have a longer time course than the rapid visual processing indexed by the P1 component (Wood and Rünger, 2016), it seemed likely that later ERP components would be more likely to reflect the operation of goal-directed processes and thus would show an effect of outcome devaluation. However, because this is the first study to use ERPs to investigate devaluation, we could not anticipate the spatiotemporal distribution of any such effect. For this reason, we analyzed the ERP data from 200 to 800 ms using mass univariate analyses, which are well suited for exploratory analysis or for delineating effect boundaries in situations with little (or no) guidance from previous studies (Groppe et al., 2011a, 2011b; Manly, 1997). Specifically, the effect of stimulus value (S^{high} vs S^{low}) was analyzed for the 200–800 ms time window (150 samples, 1 each 4 ms, as a consequence of the 250 Hz sample rate) and for all 10 midline electrodes. Nonparametric (bootstrapped) *t* tests (5000 permutations) were used for these comparisons using the EEGLAB "statcond" function (Delorme and Makeig, 2004, RRID:SCR_007292). The false discovery rate (FDR) procedure (Benjamini and Yekutieli, 2001) was used as correction for multiple comparisons across the combination of channels (10) and time points (150).

An α level of 0.05 was used to determine significance. Where the assumption of sphericity was violated in repeated-measures ANOVA, Greenhouse–Geisser-corrected *p*-values are reported, along with the corrected degrees of freedom. Relevant null results were assessed using Bayesian analysis conducted using JASP version 0.7.5.5 software (www.jasp-stats.org) and the default Cauchy prior 0.707.

Results

Behavioral results

Figure 1*C* shows behavioral data relating to participants' response choices during the reinforcement learning task. Trials in which participants responded before the stimuli appeared or failed to respond before the trial timed out (2% and 0.5% of all trials, respectively) were discarded. Accuracy on the remaining trials was calculated as the proportion of trials on which participants made a response that earned a diamond (outcome) regardless of the current value of that outcome (or equivalently, the proportion of trials on which they did not make a response that produced "Error!" feedback).

Response accuracy and response time were analyzed in separate 2 (stimulus value: S^{high}, S^{low}) × 3 (outcome devaluation: No Dev, Dev O^{high}, Dev O^{low}) repeated-measures ANOVAs. The analysis of accuracy data showed a main effect of outcome devaluation, $F_{(2,46)} = 6.83$, $p = 0.003$, $\eta_p^2 = 0.23$. Follow-up pairwise tests revealed poorer accuracy on devaluation blocks compared with no-devaluation blocks regardless of which S–R relationship was devalued(No Dev vs Dev O^{high}: $F_{(1,23)} = 12.33$, $p = 0.002$, $\eta_p^2 = 0.35$; No Dev vs Dev O^{low}: $F_{(1,23)} = 13.42$, $p = 0.001$, $\eta_p^2 = 0.37$; Dev O^{high} vs Dev O^{low}: $F < 1$. There was also a stimulus value × outcome devaluation interaction, $F_{(1.42,32.56)} = 4.99$, $p = 0.022$, $\eta_p^2 = 0.18$. This reflects the finding that accuracy was poorer in S^{high} trials compared with S^{low} trials in Dev O^{high} blocks, whereas this difference was reversed in No Dev and Dev O^{low} blocks (Fig. 1*C*). Follow-up analysis showed trends in the simple effects of stimulus value compatible with this interpretation: for No Dev blocks, $F_{(1,23)} = 3.70$, $p = 0.067$, $\eta_p^2 = 0.14$; for Dev O^{high} blocks, $F_{(1,23)} = 3.72$, $p = 0.066$, $\eta_p^2 = 0.14$; for Dev O^{low} blocks, $F_{(1,23)} = 4.63$, $p = 0.042$, $\eta_p^2 = 0.17$. These results suggest that, in No Dev blocks, participants were more focused on earning O^{high} than O^{low} (thus the more accurate responses to S^{high} in No Dev blocks). However, when one of the outcomes was devalued, participants focused on earning the outcome that was not devalued (thus the interaction). This influence of devaluation on overt behavior implicates a goal-directed system. Figure 1*C* also shows that, in devalued blocks, participants still made a "correct" response on the vast majority (>90%) of trials that presented the stimulus in which an outcome had been devalued; that is, participants were more likely to produce the response that yielded a devalued diamond than a response that produced "Error!" feedback, suggesting that they were in part motivated to avoid this negative feedback. This is a benefit of the use of negative feedback because any existing S–R links could otherwise have weakened during the devalued blocks if the pattern of responses had changed substantially (Balleine and O'Doherty, 2010).

Response time was defined as the time taken to make a response after the onset of the response options (aliens) on each trial, which occurred 800 ms after the onset of the stimulus (cookie; Fig. 1*A*). The analysis of response times did not yield any significant results ($F \leq 1.12$, $p \geq 0.306$). This may seem inconsistent with previous research showing faster execution of responses that produced high reward than low reward (MacLean and Giesbrecht, 2015); however, those previous studies used speeded response tasks, thus prioritizing rapid responses. In con-

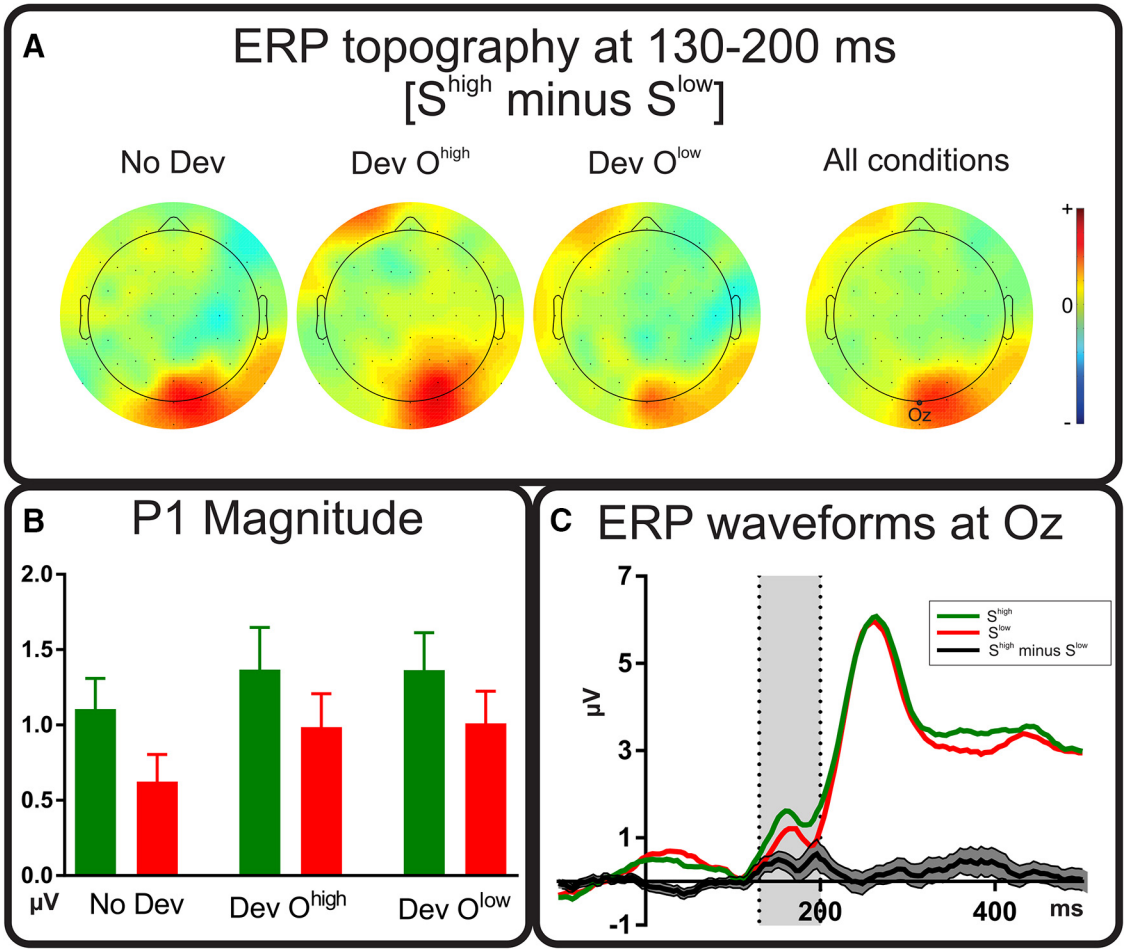

Figure 2. P1 results. ***A***, Topographical figures mapping the differences between stimuli that were more frequently related to the high-value outcome (S^{high}) and to the low-value outcome (S^{low}) for all conditions. Relative scale (minimum/maximum values): −1/+1 μV. ***B***, Magnitude of the P1 components for all conditions calculated as the mean activity at electrode Oz in the 130–200 ms time window. Error bars indicate within-participant SEM. ***C***, Mean evoked activity locked to stimulus onset (100 ms baseline also shown) for the S^{high} and S^{low} at electrode Oz. Data are the result of averaging the three outcome devaluation conditions. Black lines represent the S^{high} minus S^{low} difference waveform. Shaded areas represent the SEM of this difference.

trast, responses were unspeeded in the current task, which emphasized accuracy over speed. Therefore, it is unsurprising that effects should manifest primarily in data on response choice rather than speed. Indeed, previous research using an unspeeded response task also found no effect of reward value on response time (Luque et al., 2015).

Participants' responses on consumption trials (Fig. 1*B*) were sensitive to the outcome devaluation manipulation. Figure 1*C* shows the proportion of consumption trials on which participants selected O^{high}. This O^{high} was the more valuable option during the No Dev and Dev O^{low} blocks [when it was worth 100 points vs O^{low}, which was worth 1 point (No Dev) or zero points (Dev O^{low})], but was the less valuable option in Dev O^{high} blocks (when O^{high} was worth zero points and O^{low} was worth 1 point). ANOVA showed an effect of outcome devaluation in the distribution of responses during the consumption trials ($F_{(1.04,23.94)} = 125.7$, $p < 0.001$, $\eta_p^2 = 0.85$). As expected, the proportion of O^{high} choices was considerably lower in the Dev O^{high} blocks than in either the No Dev or the Dev O^{low} blocks (No Dev vs Dev O^{high}: $F_{(1,23)} = 141.4$, $p < 0.001$, $\eta_p^2 = 0.86$; No Dev vs Dev O^{low}: $F_{(1,23)} = 2.23$, $p = 0.149$, $\eta_p^2 = 0.09$; Dev O^{high} vs Dev O^{low}: $F_{(1,23)} = 116.02$, $p < 0.001$, $\eta_p^2 = 0.83$. Once again, these findings suggest that participants' choices on consumption trials were sensitive to the current value of the outcome, demonstrating goal-directed behavior.

ERP results

P1 time window

Figure 2*A* shows topographical maps of differences between EEG elicited by S^{high} and S^{low} (S^{high} minus S^{low}) in each of the three outcome devaluation conditions separately and averaged over the three conditions. The main difference in these maps relates to an occipital positivity, corresponding to the visual P1 ERP component. ERP waveforms at electrode Oz confirm that the differences are around the peak of the P1 component (Fig. 2*C*). Figure 2*B* shows mean P1 magnitudes (i.e., mean activity over the 130–200 ms time window). A 2 (stimulus value: S^{high}, S^{low}) × 3 (outcome devaluation: No Dev, Dev O^{high}, Dev O^{low}) ANOVA on these data revealed a main effect of stimulus value ($F_{(1,23)} = 5.70$, $p = 0.026$, $\eta_p^2 = 0.20$), with S^{high} eliciting a larger P1 than S^{low}. There was also a main effect of outcome devaluation ($F_{(2,46)} = 3.64$, $p = 0.034$, $\eta_p^2 = 0.14$). Follow-up analysis indicated that, overall, P1 was smaller in the No Dev condition than the two devaluation conditions (No Dev vs Dev O^{high}, $F_{(1,23)} = 5.00$, $p = 0.035$, $\eta_p^2 = 0.18$; No Dev vs Dev O^{low}, $F_{(1,23)} = 4.45$, $p = 0.046$, $\eta_p^2 = 0.16$; Dev O^{high} vs Dev O^{low}, $F < 1$). This result most likely reflects greater overall attention during devaluation blocks, which would be required because outcome values were not shown on each trial in these blocks, making the task more demanding. Consistent with this view, participants' response accuracy was poorer in devaluation blocks than in No Dev blocks

Table 2. BF_{10} for possible models accounting for the P1 magnitude data

	Model			
	Stimulus value	Outcome devaluation	Stimulus value + outcome devaluation	Stimulus value × outcome devaluation
BF_{10}	12.5	0.56	7.69	0.13
Interpretation	Strong support for the alternative hypothesis	Null and alternative equally supported	Substantial evidence for the alternative hypothesis	Substantial evidence for the null hypothesis

Interpretations of BF_{10} values follow the criteria in Wetzels et al. (2011).

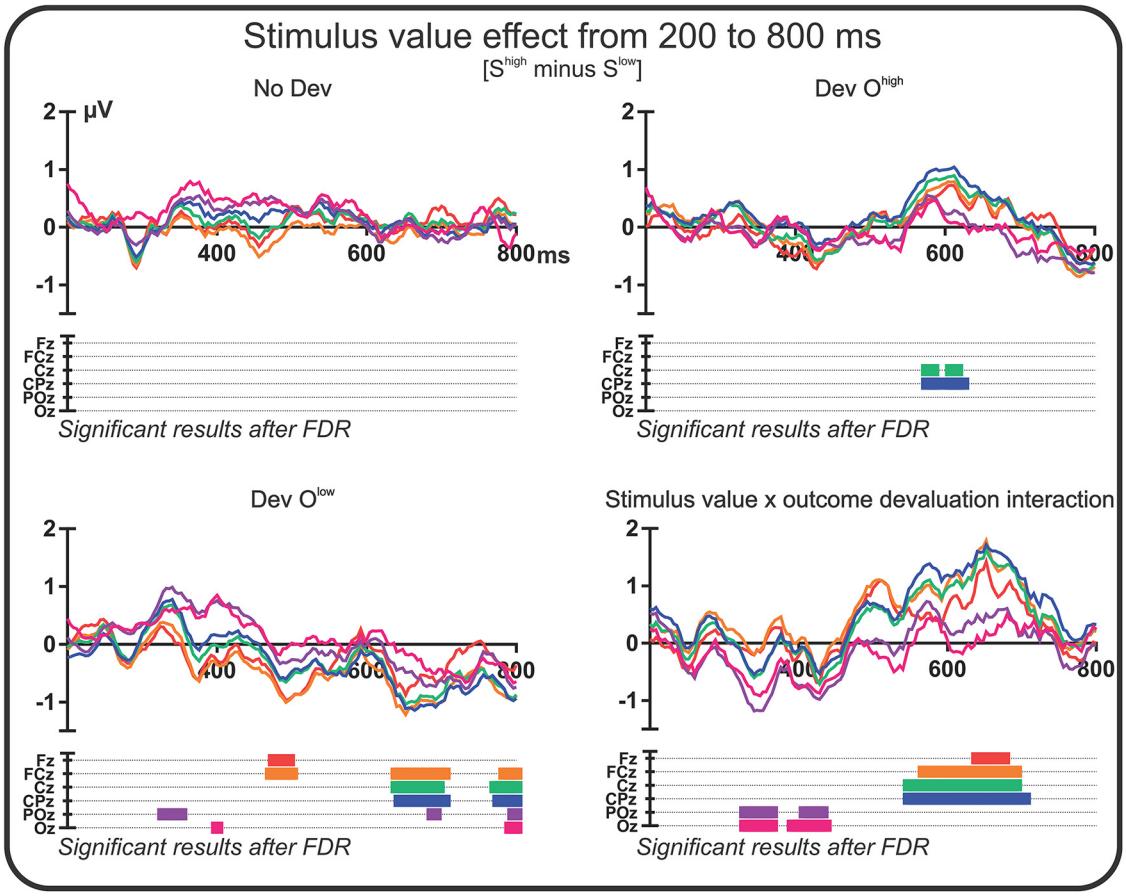

Figure 3. Long-latency ERPs. Mass univariate analysis results. Colored lines represent S^{high} minus S^{low} difference waveforms for six midline electrodes (anterior to posterior: Fz, FCz, Cz, CPz, POz, and Oz). Colored bars below the plots represent time periods during which the S^{high} minus S^{low} difference is significantly different from zero (bootstrapped *t* test, $p < 0.05$, FDR corrected). Data are shown for the No Dev condition, Dev O^{high} condition, and Dev O^{low} condition. The stimulus value × outcome devaluation interaction panel depicts the subtraction of the difference waveforms for the Dev O^{high} and O^{low} conditions.

(Fig. 1*C*). The stimulus value × outcome devaluation interaction was not significant ($F < 1$).

For present purposes, this analysis provides two key results: (1) P1 magnitude was greater for S^{high} than S^{low}, demonstrating a reward-related P1 effect, and (2) the size of this reward-related P1 effect was not influenced by the devaluation manipulation (indicated by the nonsignificant interaction). To quantify the extent to which our data favor the latter null hypotheses, we conducted a 2 × 3 Bayesian repeated-measures ANOVA on data relating to P1 magnitude. A $BF_{10} > 3$ is usually considered to reflect substantial support in favor of the alternative hypothesis and values >10 reflect strong support. Conversely, values <1/3 are considered substantial evidence and values <1/10 strong support for the null hypothesis (Wetzels et al., 2011). Results of the Bayesian analysis are shown in Table 2. As can be seen, the strongest support for the alternative hypothesis comes from the model with only a main effect of stimulus value ($BF_{10} = 12.5$). In contrast, the model based on a stimulus value × outcome devaluation interaction is a poor fit to the data, with substantial support for the null hypothesis ($BF_{10} = 0.13$).

Long-latency ERPs

Exploratory mass univariate analysis was conducted to assess possible goal-directed brain activity at longer latencies. Figure 3 shows results from this analysis using nonparametric bootstrapping tests to compare activity elicited by S^{high} and S^{low} stimuli (stimulus value effects) across the long-latency time window (200–800 ms) of their presentation (FDR corrected). In the Dev O^{high} condition, the stimulus value effect was significant for several time points in the 570–630 ms range, with a more positive EEG signal for the S^{high} stimulus than for the S^{low} stimulus. This difference was mirrored in the Dev O^{low} condition, in which (toward the end of the long-latency time window) activity elicited by the S^{high} stimulus was less positive than that elicited by the S^{low} stimulus. This trend continued until the offset of the stimulus (reaching significance for several time points from 640–700 and

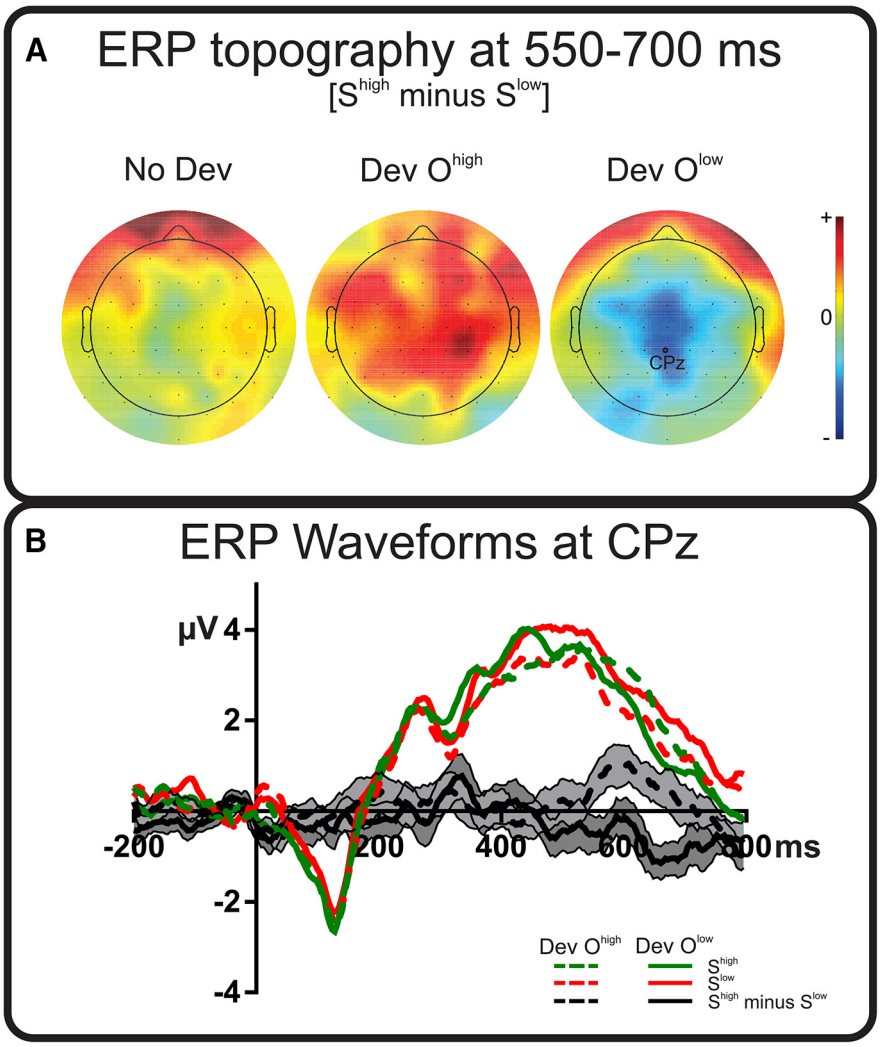

Figure 4. Long-latency ERPs. Results in the 550–700 ms time window are shown. ***A***, Topographical figures mapping the differences between stimuli that were more frequently related to high-value outcome (S^{high}) and to low-value outcome (S^{low}) for all conditions. Relative scale (minimum/maximum values): $-1/+1$ μV. ***B***, Mean evoked activity locked to stimulus onset (100 ms baseline also shown) for the S^{high} and S^{low} at relevant electrode CPz. Data for the two devalued conditions, Dev O^{high} and Dev O^{low}, are shown. Black lines represent the S^{high} minus S^{low} difference waveforms for each devaluation condition. Shaded areas represent SEM of these differences.

770–800 ms) in a widely spread distribution including central and parietal sites. Therefore, at late time points (570–800 ms), the pattern of activity elicited by the same stimuli (S^{high} and S^{low}) depended critically on the current value of the outcome with which they were associated. Phrased differently, in both Dev O^{high} and Dev O^{low} blocks, greater activity was elicited by the stimulus that signaled the devalued outcome (S^{high} in Dev O^{high} blocks; S^{low} in Dev O^{low} blocks) than by the stimulus that signaled the unchanged value outcome (S^{low} in Dev O^{high} blocks; S^{high} in Dev O^{low} blocks). This sensitivity of brain activity to the current value of predicted outcomes is consistent with goal-directed processing of the stimulus. Importantly, there were no significant stimulus value (S^{high} vs S^{low}) effects in the No Dev condition, confirming that the outcome devaluation effects just described were not present in the absence of devaluation (see Discussion for further implications of this result).

To compare more directly reward effects on brain activity in the different devaluation conditions, we submitted the stimulus value × outcome devaluation interaction directly to a mass univariate analysis. Figure 3 shows the spatiotemporal dynamics of this interaction, generated by subtracting difference waveforms for the Dev O^{low} condition from those for the Dev O^{high} condition (so that positive values in this interaction waveform indicate that the S^{high} minus S^{low} difference was larger in the Dev O^{high} condition and vice versa). Significant differences between devaluation conditions were observed at occipital sites 330–440 ms from stimulus onset and these differences became more accentuated with a more widespread distribution (including central and parietal sites) from 550–700 ms, confirming previous analyses. Figure 4 shows topographical maps of S^{high} minus S^{low} differences during this period (Fig. 4*A*), along with ERP waveforms elicited by S^{high} and S^{low} stimuli during the whole stimulus duration (Fig. 4*B*). Consistent with the argument advanced above, these findings support the idea that brain activity elicited by S^{high} and S^{low} in the 550–700 ms time window depended critically and significantly on the devaluation condition; that is, on the current values of the outcomes signaled by these stimuli. The implication is that this interaction reflects the operation of goal-directed brain processing.

Discussion

An influential idea concerning attention is that perceptual prioritization can be a form of habitual response (Anderson, 2016). It is well established that early components of stimulus-locked ERPs (i.e., P1 component) reflect stimulus salience computed by the extrastriate visual cortex (Heinze et al., 1994) and prior studies have shown increased P1 for those stimuli reliably related with high-value reward (Hickey et al., 2010; MacLean and Giesbrecht, 2015). Here, we provide the first evidence to support this assumption by testing whether reward-related P1 is resistant to transient outcome devaluations, a cardinal feature of habitual responses. Therefore, the present experiment is the first to test strictly the existence of an "attentional habit" as evident in the P1 component.

Our results revealed that the P1 component elicited by a reward-related stimulus was affected by the magnitude of its more frequent outcome; that is, S^{high} elicited a P1 component that was larger (more positive) than that elicited by S^{low}. This result indicates that reward learning produced neural changes in participants' visual cortex, effectively increasing their perceptual prioritization of S^{high}. Importantly, this reward-related perceptual prioritization effect was not affected by transient variations in the value of the outcomes associated with these stimuli, as expected for habitual responses.

Previous studies have reported that P1 is sensitive to reward value even when the analyzed stimuli were no longer predictive of reward (Hickey et al., 2010; MacLean and Giesbrecht, 2015). Such studies tested responses during an extinction (no reinforcement) phase. Notably, although responses to the critical stimuli

were no longer rewarded in this extinction phase, the value of the rewards that had been associated with those stimuli did not change. Therefore, previous literature did not assess directly the influence of outcome value over perceptual prioritization, which requires changing outcome values in one stimulus while holding the other constant. Outcome value manipulation is important because, even when the outcomes are no longer available, the goal-directed system would presumably maintain the neural representation for outcome values (Ostlund and Balleine, 2007). Therefore, in these previous studies, reward-related P1 effects could still reflect the actual outcome value established and maintained by the goal-directed system without requiring the participation of the habit system. The same reasoning does not apply to the current data; here, outcome values were modulated transiently via devaluation instructions and overt behavior (responses on reinforcement learning trials and consumption trials) showed that participants had updated their knowledge accordingly. The persistence of reward-related P1 effects despite these transient changes in outcome value therefore represents the strongest evidence to date of habit-like neural activity in the human visual cortex.

Analysis of longer-latency stimulus-locked ERP activity showed that neural activity was sensitive to the outcome devaluation manipulation in two spatiotemporal regions, first at occipital sites from 330–440 ms and then in a wider distribution, with maximum effects around centroparietal sites in a 550–700 ms time window. Importantly, the latter effect was produced because the EEG signal was more positive for those stimuli that were specifically associated with devalued outcomes. That is, significant stimulus value effects were observed in both Dev O^{high} and Dev O^{low} conditions, but in different directions. This effect reflects the pattern of brain activity expected from a goal-directed system. Interestingly, this differential brain activity did not merely reflect the value of the anticipated outcome because no significant difference between S^{high} and S^{low} was observed in the No Dev condition despite the difference in outcome values in these blocks.

Given its spatiotemporal features, the goal-directed effect observed in the devaluation conditions is most likely due to changes in the P3b, a component that has also been related to executive functions such as working memory (Polich, 2007; but see Verleger, 2008). Notably, the loop between prefrontal cortex and the basal ganglia (caudate nucleus), which supports goal-directed actions, has also been related to working memory functions (Sharp et al., 2016). Therefore, a reasonable hypothesis is that this goal-directed modulation of P3b reflects working memory activity, retrieving and processing changes in outcome value on each trial (see also Otto et al., 2013). Specifically, in No Dev blocks, the outcomes associated with both stimuli (S^{high} and S^{low}) took their most frequent, "standard" values (+100 and +1, respectively). Working memory demands in retrieving these standard values would be similar, as reflected in the similar P3b to S^{high} and S^{low} in No Dev blocks. In contrast, in devaluation blocks, one of the outcomes paired with one of the stimuli took on an unusual (zero) value. Greater working memory resources would be required on these trials to retrieve this temporary updated value and/or suppress the standard value of the outcome associated with this stimulus, as reflected in the greater P3b elicited by the stimulus paired with the devalued outcome in each type of devaluation block.

It is well established that behavioral control by the goal-directed system dominates early in training, whereas the habit system dominates with extended training (Balleine and O'Doherty, 2010). Given that our participants did not undertake massive training (which usually entails training across several sessions and days; Tricomi et al., 2009), their actions were probably goal directed. Our results are consistent with this interpretation: in particular, on consumption trials, participants were much more likely to select the outcome that had the higher current value and the responses on the reinforcement learning trials suggest that participants were more focused on earning the outcome that currently had higher value (Fig. 1*C*). Therefore, it seems that participants' attentional habit system was active during the experiment (i.e., it was somehow modulating activity in participants' visual cortex) even when their behavior was goal directed. Our results are thus consistent with dual-process theories of learning, which assume that both goal-directed and habit systems operate in parallel. We found distinct habitual and goal-directed ERP components over each trial in the early and late stages, respectively. This pattern of evidence supports models assuming that the final behavioral output is a product of both systems or arbitration occurs at a late stage of response selection (Balleine and O'Doherty, 2010; Lee et al., 2014).

The results of the present study also help to establish the temporal order in which habit and goal-directed systems operate. Habit learning and behavior are sustained by the so-called sensorimotor loop, which connects the sensorimotor and motor cortex with the medial and posterior putamen (Yin and Knowlton, 2006). In an fMRI study with humans, Tricomi et al. (2009) found increased BOLD activity in the posterior putamen over the course of individual practice sessions and across days of practice. Therefore, in our task, the putamen might be activated by the presentation of the stimulus, producing progressive strengthening of S–R links. It is difficult to detect putamen activity using EEG because it is a subcortical region (Luck, 2014), so we cannot determine whether the habit system itself was activated first, at the same time, or after the goal-directed system. However, we can establish that a habit-like signal did influence the first stages of stimulus processing as early as 130 ms after stimulus presentation. In addition, our EEG data revealed no goal-directed activity until 330 ms after stimulus presentation. Therefore, our data suggest that faster habit-like processing took place before goal-directed processing. This order of information processing maps well onto associative theories claiming that S–R activation is faster than goal-directed processing (e.g., the associative-cybernetic model; Dickinson and Balleine, 1994). Interestingly, similar predictions are made by dual-process theories of decision making and reasoning (Evans, 2008). In these theories, fast automatic mechanisms operate first by default and without requiring significant cognitive resources. If, however, a conflict-monitoring system detects the need for additional cognitive control, a slower and more cognitively demanding mechanism would take control of behavior (Kahneman, 2011).

The current study raises important questions that could be addressed in future experiments. For instance, it is well known that Pavlovian conditioned responses, which are learned from direct pairings between a conditioned stimulus (e.g., a tone) and an unconditioned stimulus (e.g., food), play an important role in reinforcement learning (de Wit and Dickinson, 2009). Future research might investigate to what extent the habit-like P1 effect observed here is the result of learning different Pavlovian values for the S^{high} and S^{low}. In addition, future studies could extend the generality of the current findings by using an R–O contingency degradation strategy (Dickinson et al., 1998) to further assess the habitual nature of the reward-related P1. This test of habits manipulates the effectiveness of the response in producing outcomes (rather than manipulating the value of the outcomes as in the current study).

In summary, the present results provide for the first time evidence of rapid, habit-like activity in visual cortex (P1 component), which is followed by a slower, goal-directed brain activity

as shown in the P3b component. These results highlight the importance of attentional/perceptual processing for habit learning and vice versa; are compatible with the idea that habit and goal-directed systems are activated in parallel even during early stages of training; and suggest that, during reinforcement learning, stimulus processing is prioritized by habit-like attentional processes and subsequently by goal-directed processes that adapt the stimulus value through working memory.

References

Anderson BA (2016) The attention habit: how reward learning shapes attentional selection. Ann N Y Acad Sci 1369:24–39. CrossRef Medline

Anderson BA, Kuwabara H, Wong DF, Gean EG, Rahmim A, Brašić JR, George N, Frolov B, Courtney SM, Yantis S (2016) The role of dopamine in value-based attentional orienting. Curr Biol 26:550–555. CrossRef Medline

Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35:48–69. CrossRef Medline

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 1165–1188.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21. CrossRef Medline

de Wit S, Dickinson A (2009) Associative theories of goal-directed behaviour: a case for animal-human translational models. Psychol Res 73:463–476. CrossRef Medline

de Wit S, Corlett PR, Aitken MR, Dickinson A, Fletcher PC (2009) Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. J Neurosci 29:11330–11338. CrossRef Medline

Dickinson A, Balleine B (1994) Motivational control of goal-directed action. Animal Learning & Behavior 22:1–18. CrossRef

Dickinson A, Squire S, Varga Z, Smith JW (1998) Omission learning after instrumental pretraining. Q J Exp Psychol 51:271–286.

Di Russo F, Martínez A, Sereno MI, Pitzalis S, Hillyard SA (2002) Cortical sources of the early components of the visual evoked potential. Hum Brain Mapp 15:95–111. CrossRef Medline

Evans JS (2008) Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol 59:255–278. CrossRef Medline

Gillan CM, Morein-Zamir S, Urcelay GP, Sule A, Voon V, Apergis-Schoute AM, Fineberg NA, Sahakian BJ, Robbins TW (2014) Enhanced avoidance habits in obsessive-compulsive disorder. Biol Psychiatry 75:631–638. CrossRef Medline

Gillan CM, Otto AR, Phelps EA, Daw ND (2015) Model-based learning protects against forming habits. Cogn Affect Behav Neurosci 15:523–536. CrossRef Medline

Groppe DM, Urbach TP, Kutas M (2011a) Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology 48:1711–1725. CrossRef Medline

Groppe DM, Urbach TP, Kutas M (2011b) Mass univariate analysis of event-related brain potentials/fields II: simulation studies. Psychophysiology 48:1726–1737. CrossRef Medline

Heinze HJ, Luck SJ, Munte TF, Gös A, Mangun GR, Hillyard SA (1994) Attention to adjacent and separate positions in space: An electrophysiological analysis. Perception & Psychophysics 56:42–52. CrossRef

Hickey C, Chelazzi L, Theeuwes J (2010) Reward changes salience in human vision via the anterior cingulate. J Neurosci 30:11096–11103. CrossRef Medline

Hillyard SA, Vogel EK, Luck SJ (1998) Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philos Trans R Soc Lond B Biol Sci 353:1257–1270. CrossRef Medline

Kahneman D (2011) Thinking, fast and slow. London: MacMillan.

Kleiner M, Brainard DH, Pelli DG (2007) What's new in Psychtoolbox-3? Perception 36:1–16.

Lee SW, Shimojo S, O'Doherty JP (2014) Neural computations underlying arbitration between model-based and model-free learning. Neuron 81:687–699. CrossRef Medline

Le Pelley ME, Mitchell CJ, Beesley T, George DN, Wills AJ (2016) Attention and associative learning in humans: an integrative review. Psychol Bull 142:1111–1140. CrossRef Medline

Luck SJ (2014) An introduction to the event-related potential technique. Cambridge, MA: MIT.

Luque D, Morís J, Rushby JA, Le Pelley ME (2015) Goal-directed EEG activity evoked by discriminative stimuli in reinforcement learning. Psychophysiology 52:238–248. CrossRef Medline

MacLean MH, Giesbrecht B (2015) Neural evidence reveals the rapid effects of reward history on selective attention. Brain Res 1606:86–94. CrossRef Medline

Manly BFJ (1997) Randomization, bootstrap, and Monte Carlo methods in biology, ed 2. London: Chapman and Hall.

Ostlund SB, Balline BW (2007) Selective reinstatement of instrumental performance depends on the discriminative stimulus properties of the mediating outcome. Learn Behav 35:43–52. CrossRef Medline

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013) Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A 110:20941–20946. CrossRef Medline

Pearson D, Osborn R, Whitford TJ, Failing M, Theeuwes J, Le Pelley ME (2016) Value-modulated oculomotor capture by task-irrelevant stimuli is a consequence of early competition on the saccade map. Atten Percept Psychophys 78:2226–2240. CrossRef Medline

Polich J (2007) Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 118:2128–2148. CrossRef Medline

Sharp ME, Foerde K, Daw ND, Shohamy D (2016) Dopamine selectively remediates "model-based" reward learning: a computational approach. Brain 139:355–364. CrossRef Medline

Smith NK, Cacioppo JT, Larsen JT, Chartrand TL (2003) May I have your attention, please: electrocortical responses to positive and negative stimuli. Neuropsychologia 41:171–183. CrossRef Medline

Tricomi E, Balleine BW, O'Doherty JP (2009) A specific role for posterior dorsolateral striatum in human habit learning. Eur J Neurosci 29:2225–2232. CrossRef Medline

Verleger R (2008) P3b: Towards some decision about memory. Clin Neurophysiol 119:968–970. CrossRef Medline

Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ (2011) Statistical evidence in experimental psychology an empirical comparison using 855 t tests. Perspect Psychol Sci 6:291–298. CrossRef Medline

Wood W, Rünger D (2016) Psychology of habit. Annu Rev Psychol 67:289–314. CrossRef Medline

Yin HH, Knowlton BJ (2006) The role of the basal ganglia in habit formation. Nat Rev Neurosci 7:464–776. CrossRef Medline