Cogn Affect Behav Neurosci (2017) 17:642–651
DOI 10.3758/s13415-017-0502-3

Reward-based contextual learning supported by anterior cingulate cortex

Akina Umemoto[1,2] · Azadeh HajiHosseini[1] · Michael E. Yates[1] · Clay B. Holroyd[1]

Published online: 24 February 2017

Abstract The anterior cingulate cortex (ACC) is commonly associated with cognitive control and decision making, but its specific function is highly debated. To explore a recent theory that the ACC learns the reward values of task contexts (Holroyd & McClure in *Psychological Review, 122,* 54–83, 2015; Holroyd & Yeung in *Trends in Cognitive Sciences, 16,* 122–128, 2012), we recorded the event-related brain potentials (ERPs) from participants as they played a novel gambling task. The participants were first required to select from among three games in one "virtual casino," and subsequently they were required to select from among three different games in a different virtual casino; unbeknownst to them, the payoffs for the games were higher in one casino than in the other. Analysis of the reward positivity, an ERP component believed to reflect reward-related signals carried to the ACC by the midbrain dopamine system, revealed that the ACC is sensitive to differences in the reward values associated with both the casinos and the games inside the casinos, indicating that participants learned the values of the contexts in which rewards were delivered. These results highlight the importance of the ACC in learning the reward values of task contexts in order to guide action selection.

Keywords Contextual learning · Reinforcement learning · Anterior cingulate cortex · ERP · Dopamine · Learning reward value

Anterior cingulate cortex (ACC) is involved in decision making and cognitive control, but its exact role in this domain is debated. In particular, it has been proposed that ACC instigates immediate changes to behavior following experienced response conflict (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Kerns et al., 2004), and that ACC is involved in value-guided decision making (Holroyd & Coles, 2002; Kolling et al., 2016; Shenhav, Cohen, & Botvinick, 2016). Although value-based theories disagree on the specifics, they converge on the suggestion that the ACC modulates behavioral policies over multiple episodes on the basis of a cost–benefit analysis of rewards and penalties (Ebitz & Hayden, 2016).

We have previously proposed that ACC utilizes reward prediction error signals (RPEs) carried by the midbrain dopamine system for the purpose of reinforcing adaptive behaviors, and that this process is revealed by a component of the event-related brain potential (ERP) called the *reward positivity* (Holroyd & Coles, 2002). More commonly known as the *feedback error-related negativity*, this component was originally associated with a negative deflection in the ERP that is elicited by negative performance feedback, but recent evidence has indicated that the ERPs to both positive and negative feedback are driven by the positive outcome (Holroyd, Pakzad-Vaezi, & Krigolson, 2008; Proudfit, 2015). Consistent with the theory, the reward positivity appears to be generated in ACC (Becker, Nitsch, Miltner, & Straube, 2014; Miltner, Braun, & Coles, 1997), and a wealth of evidence has indicated

Electronic supplementary material The online version of this article (doi:10.3758/s13415-017-0502-3) contains supplementary material, which is available to authorized users.

✉ Akina Umemoto
akumemoto@gmail.com

[1] Department of Psychology, University of Victoria, Victoria, British Columbia, Canada

[2] Present address: Department of Psychiatry and Neurosciences, Institute of Biomedical and Health Sciences, Hiroshima University, 1-2-3, Kasumi, Minami-ku, Hiroshima City, Hiroshima 734-8553, Japan

that it indexes an RPE (Sambrook & Goslin, 2015; Walsh & Anderson, 2012). By contrast, the evidence that the reward positivity is associated with a reinforcement-learning (RL) process is less clear (for reviews, see Holroyd & Umemoto, 2016; Walsh & Anderson, 2012). For example, when participants are given instructions that result in an immediate change in task performance, the trial-to-trial changes in reward positivity amplitude continue to reflect a slow learning process (Walsh & Anderson, 2011).

Nevertheless, it should be noted that the original RL account of the reward positivity does not, in fact, hold the ACC responsible for instigating such trial-to-trial changes in task performance. Rather, the model proposes that ACC learns the value of entire task policies and selects the policies on the basis of those learned values (Holroyd & Coles, 2002; cf. Daw, Niv, & Dayan, 2005). Recent developments of this idea have suggested that ACC encodes the value of the task context itself, rather than the value of the individual actions carried out within the task (Holroyd & McClure, 2015; Holroyd & Yeung, 2011, 2012). Thus, for example, the reward positivity is elicited even when participants passively view reward stimuli in the absence of overt behavior, suggesting that "the reward signal…is also used by ACC to evaluate more distal events and more general action plans that are not directly task-related" (Yeung, Holroyd, & Cohen, 2005, p. 542). In turn, these contextual values are utilized by ACC to motivate task performance (Umemoto & Holroyd, 2016).

To investigate this issue, we recorded the reward positivity from participants engaged in a novel casino-gambling task that allowed for learning the reward value of the task context (cf. Diuk, Tsai, Wallis, Botvinick, & Niv, 2013). Participants selected between different games in two different virtual casinos in order to earn money. Unbeknownst to them, the overall payoff in one casino was better than the overall payoff in the other casino. We asked whether the reward positivity would reveal that participants had learned not just the values of individual games in the casinos, but also the values of the casinos that housed the games—even though this higher-order information was irrelevant to task performance. Importantly, because the participants were allowed to enter each casino only once, the task design prevented participants from learning the casino values on the basis of simple stimulus–feedback contingencies from trial to trial. Instead, the task encouraged participants to learn values for internal representations of the two casino contexts (Holroyd & McClure, 2015).

Method

Participants

A total of 27 undergraduate students were recruited from the University of Victoria Department of Psychology subject pool either to fulfill a course requirement or earn bonus credits. The number of participants was determined on the basis of past reward positivity studies (~15 participants); because we were employing a novel task design, the sample size was doubled to approximately 30 participants total. The data of two participants were excluded due to self-reported neurological disorders (aneurysm and concussion). The remaining 25 participants (20 females, five males; three left-handed; age range = 18–28 years, mean age = 21.2 ± 2.5 years) all reported normal or corrected-to-normal vision. Each participant received a monetary bonus that depended on task performance (see below). All participants provided informed consent as approved by the local research ethics committee. The experiment was conducted in accordance with the ethical standards prescribed in the 1964 Declaration of Helsinki.

Task, design, and procedure

Participants were seated comfortably in front of an LCD computer monitor (1,024 × 1,280 pixels, a 60-Hz refresh rate) at a distance of about 60 cm in an electromagnetically shielded, dimly lit room. The task was programmed in MATLAB (MathWorks, Natick, MA, USA) using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997). Participants were asked to position their hand and forearm so that their dominant hand rested below a response box (Empirisoft, Model S/N k5521-04) placed in front of them. Participants were provided with both written and verbal instructions that explained the procedure. They were told to maintain correct posture and to minimize head movements and eye blinks during the experiment.

At the outset of the experiment, participants were told that they would be provided with imaginary casino-specific tokens (180 tokens total, see below) to play a series of gambling games in two different casinos (Fig. 1a, 1st panel from left). Each token allowed for a single game play inside the casino specific to that token. Participants were told that their accumulated winnings would be given to them upon task completion, and they were encouraged to explore the games to maximize their earnings. The experiment was divided into five distinct phases, as follows: (1) practice phase, (2) deep-processing phase, (3) initial playing phase, (4) search phase, and (5) second playing phase.

1. *Practice phase*: Participants first practiced the task by using ten imaginary tokens to select and play among three games ten times as described below (initial playing phase). Following the practice, they were physically paid 25 cents, corresponding to a 50% reward probability. Note that the specific procedure for each trial in the practice phase was identical to that in the initial and second playing phases (see Points 3 and 5 below), except that two

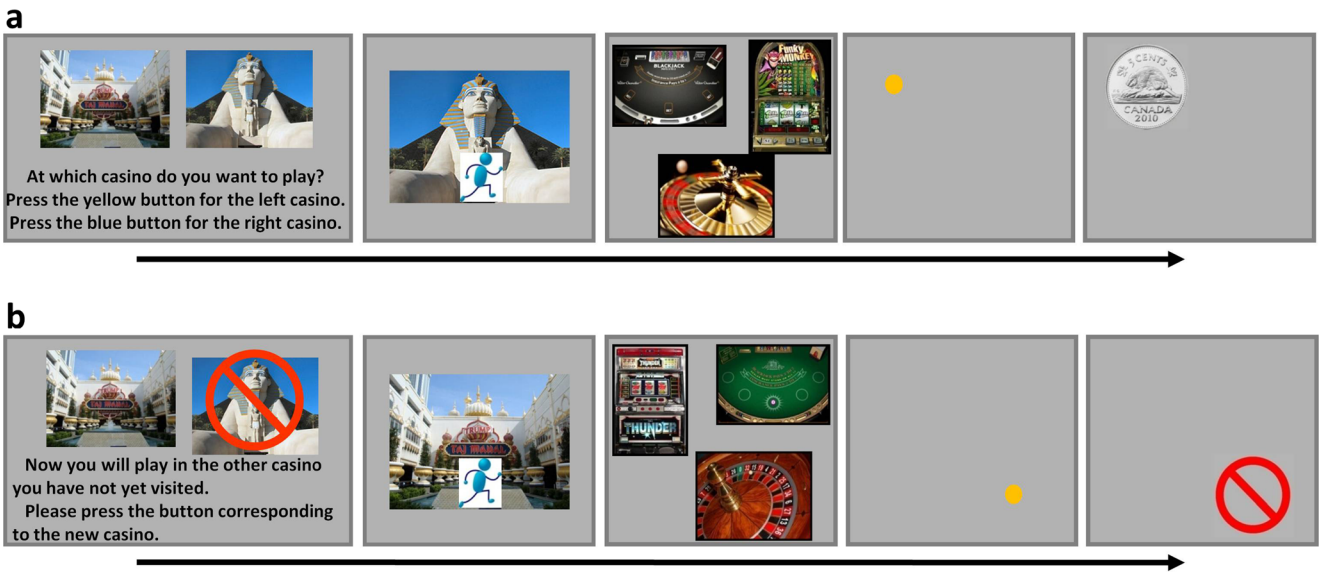

Fig. 1 Example trials during the initial playing phase. (**a**) Initial sequence of events corresponding to participant selection of the first casino (here, the "Luxor" casino), followed by an example trial in which the participant chose between three games and then either did or did not receive a five-cent reward (here, the reward). This sequence of a game choice followed by feedback repeated, for 90 trials total. The orange dot at the location of the chosen game was intended to direct the participant's gaze to the feedback location. (**b**) Subsequent sequence of events corresponding to participant selection of the casino that had not yet been visited (here, the "Taj Mahal" casino; note that the data for this study were collected before the beginning of the 2016 United States presidential campaign), followed by another 90 trials of game playing. In this example, selecting the bottom right game resulted in a no-reward outcome.

fresh sets of game images were used during the subsequent phases.

2. *Deep-processing phase*: After being familiarized with the task procedure, the participants were presented with two detailed images of the casinos, one of each, consecutively on the computer screen (the presentation order was counterbalanced across participants). To encourage deep processing of the stimuli (Craik & Lockhart, 1972), the participants were instructed: "Below are pictures of two casinos. During the experiment you will have the opportunity to play at both of these casinos. Please describe each casino in detail below their image. Provide details of their appearance, the impressions they give, their names, and so on. Please write as much as you can about them in the space provided." Participants then used the keyboard to type, in a few minutes, as many details as possible that they had observed about each casino in a half-page space on the screen below each image. This phase was intended to facilitate encoding the two casinos as two distinct and memorable contexts, given that the participants entered each casino only once.
3. *Initial playing phase*: Next, participants were told that they had been given 90 tokens to play 90 consecutive games in one casino, followed by 90 tokens to play 90 consecutive games in the other casino. They were then presented with the two casino images (12.7° × 11.5°) side-by-side on the computer screen with instructions underneath (Fig. 1a, 1st panel from left). They then selected a casino by pressing one of two corresponding buttons on a response box, after which a cartoon image of a person (2° × 2.6°) was depicted entering the selected casino (Fig. 1a, 2nd panel). Participants were then presented with an image of a roulette wheel, an image of a blackjack table, and an image of a slot machine (7.9° × 6.3° each, selected from two sets of three games randomly assigned to the two casinos and counterbalanced across participants) arranged across the display as depicted in Fig. 1a, 3rd panel. The positions of three game stimuli remained fixed throughout the experiment.

Once "inside" the casino, on each trial the participants "played" a game by pressing one of three corresponding response buttons on a response box, at which point the game images disappeared and a small orange fixation dot (0.3° × 0.3°) appeared at the chosen game location for 500 ms (Fig. 1a and b, 4th panel). Then either a silver coin (2.6° × 2.6°), representing 5 cents reward (Fig. 1a, 5th panel) or a red circle with a slash "/" over it, representing no reward (2.6° × 2.6°) (Fig. 1b, 5th panel), was presented for 800 ms in the location of the fixation dot. The next trial began immediately afterward, with the same three games being presented in the identical locations (Fig. 1a, 3rd panel). Participants played in the first casino until all 90 tokens for that casino "ran out." Next, they were again presented with the images of the two casinos side by side and were asked to recall which casino they had just visited by pressing the corresponding button.

If they failed to answer this question correctly, they were required to start the initial playing phase again from the beginning in the same casino.[1] After correctly answering the question, participants were again shown the two casino images side by side, with a red circle and slash overlain over the image of the previously selected casino, with instructions underneath (Fig. 1b, 1st panel). Participants then progressed through the task in the same manner as in the first casino (Fig. 1b, 2nd panel) by playing games for an additional 90 trials on three new machines (Fig. 1b, 3rd–5th panels). Upon finishing the 90 trials in this casino, participants were again asked to recall in which casino they had just played by pressing one of two corresponding buttons.[2] Critically, unbeknownst to the participants, the three games were associated at one casino with reward probabilities of 60%, 70%, and 80%, and at the other casino with reward probabilities of 20%, 30%, and 40%, such that the game play at one casino yielded a higher payoff than at the other. To reinforce this difference in value, on two occasions in the "good" casino the feedback consisted of an image of a gold coin worth $1 for 1 s (i.e., an image of the Canadian $1 coin, with "$1" written in a large font on the coin surface, on trial numbers 30 and 60; 2.6° × 2.6°). At the completion of this initial playing phase, participants were given a paper-and-pencil questionnaire ("Questionnaire 1") regarding which casino they preferred and why they preferred that casino, at which casino they had earned the most money, and whether they had noticed the $1 bonus. Participants were paid their accumulated winnings (across casinos) upon completion of the initial playing phase of the experiment (~$6 CAN).

4. *Search phase*: Participants were then told that they had spent all of their casino tokens and were instructed that they could search for additional tokens by searching for them throughout the city. For this purpose, the participants were presented with an overview of a small section of Las Vegas (15° × 21°) overlain with a 10 × 16 grid (Fig. 2, top row, 1st panel from left). On each trial during this phase, they were asked to select an unexplored grid square by clicking on it with a mouse (Fig. 2, 2nd panels). Immediately after a response, the chosen square turned black, which invalidated that choice for future trials. After a 500-ms delay, an orange central fixation dot (0.4° × 0.4°) appeared for 1 s (Fig. 2, 3rd panels), indicating to participants that a token stimulus would appear in that location shortly. Then one of the two casino token images (3.7° × 3.7°) appeared at the position of the fixation dot for 800 ms, indicating that the participant had "found" that token at that location (Fig. 2, 4th panels). The next trial began immediately thereafter; previously selected squares remained unavailable for selection for the remainder of the phase. Unbeknownst to participants, the tokens were distributed across the map pseudorandomly and occurred with equal probability. However, to reinforce the cover story, the participants were told that (1) each cell contained a token, (2) the token type depended on where they searched, and (3) they would be allowed to spend their accumulated tokens in the associated casinos. Participants first practiced the search task for four trials, which yielded two tokens for each casino. Immediately thereafter, they utilized these tokens by visiting and playing two games in each of the two casinos according to the procedure outlined in the initial playing phase. Following these practice trials, the participants began the actual search phase. They were not informed about how many times they would be allowed to search the city, but the task ended after 80 trials, yielding 40 tokens per casino.
5. *Second playing phase*: Finally, the participants spent their acquired tokens by playing 40 games in each casino, as promised, following the procedure described in the initial playing phase. On average, they earned about $5 CAN during this phase. Because we were interested in the ERPs from the initial playing phase, the data from the second playing phase are not reported here.

At the completion of the second playing phase, participants answered a paper-and-pencil questionnaire ("Questionnaire 2") that assessed their level of task engagement, strategies utilized, awareness of the reward probability manipulation across games, token preferences, and how important it was for participants to find the preferred token in the search phase, on a scale of 1 to 5 (1 = *not important at all* and 5 = *very important*). We analyzed the data using IBM SPSS Statistics 19. For each statistical result, we report the mean, standard deviation (*SD*), and effect size.

We predicted that a larger reward positivity would be elicited by the good casino tokens than by the bad casino tokens, indicating that participants had learned the values for the different casino contexts, despite this information being irrelevant to the task performance. Importantly, each casino was selected only once at the start of each block of 90 trials, which prevented participants from learning the casino values by way of a series of externally presented stimulus–feedback associations. Rather, the task encouraged participants to associate reward value with internally maintained representations of the casino contexts.

[1] Two participants repeated this step of the procedure. The behavioral and ERP results remained the same, irrespective of whether these data were included or excluded from the analyses.

[2] All participants ($N = 25$) correctly answered this question.

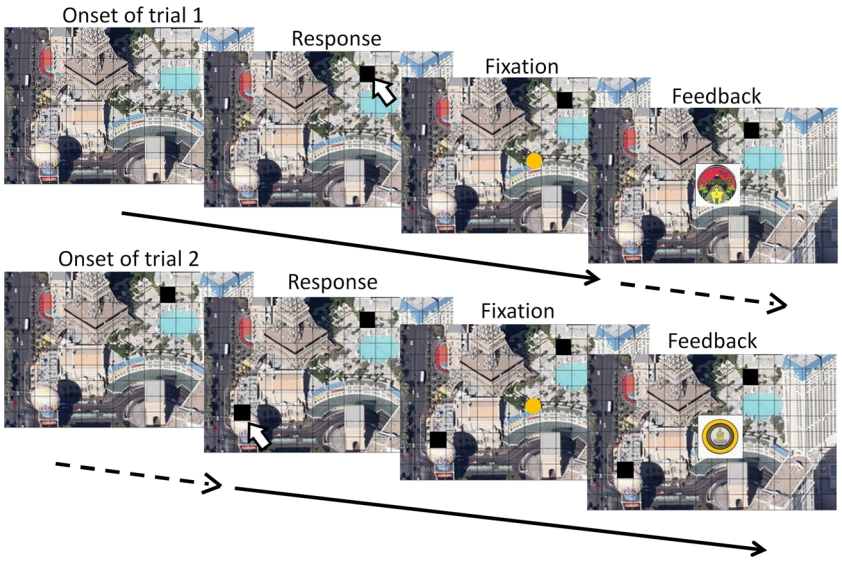

Fig. 2 Two example trials from the search phase. Participants used a mouse to move a cursor (the white arrow) and click on one grid cell at a time to look for hidden tokens. Each click on a new location revealed one token from either of the two casinos (here, a Luxor token on Trial 1 and a Taj Mahal token on Trial 2). Each previously-searched cell was blackened, indicating to participants that these cells were no longer searchable. An orange fixation dot at the center of the screen was intended to draw the participant's gaze to the feedback stimulus location. Participants searched the grid 80 times, but were not told the total number. Note that the cursor and feedback images are enlarged for the purpose of this illustration.

ERP acquisition and processing

The electroencephalogram (EEG) was recorded using a montage of 41 electrode sites in accordance with the extended International 10–20 System (Jasper, 1958). Signals were acquired using Ag/AgCl ring electrodes mounted in a nylon electrode cap with an abrasive, conductive gel (EASYCAP GmbH, Herrsching-Breitbrunn, Germany). Signals were amplified by low-noise electrode differential amplifiers with a frequency response high cutoff at 50 Hz (90-dB octave rolloff), and digitized at a rate of 250 samples per second. The digitized signals were recorded to disk using the Brain Vision Recorder software (Brain Products GmbH, Munich, Germany). The interelectrode impedances were maintained below 10 kΩ. Two electrodes were also placed on the left and right mastoids, and the EEG was recorded using the average reference. The electrooculogram (EOG) was recorded for the purpose of artifact correction; horizontal EOG was recorded from the external canthi of both eyes, and vertical EOG was recorded from the suborbit of the right eye and electrode channel Fp2.

Postprocessing and data visualization were performed using the Brain Vision Analyzer software (Brain Products GmbH). The digitized signals were filtered using a fourth-order digital Butterworth filter with a passband of 0.10–20 Hz. An 800-ms epoch of data, extending from 200 ms prior to 600 ms following the presentation of each reward feedback stimulus, was used to segment the data for waveform analysis. Ocular artifacts were corrected using an eye movement correction algorithm (Gratton, Coles, & Donchin, 1983). The EEG data were re-referenced to linked mastoid electrodes. The data were baseline-corrected by subtracting from each sample the mean voltage associated with that electrode during the 200-ms interval preceding stimulus onset. Muscular and other artifacts were removed using a ±150-μV level threshold and a ±35-μV step threshold as rejection criteria. ERPs were then created for each electrode and participant by averaging the single-trial EEG according to the reward and no-reward feedback conditions for the casino games and the casino tokens: For the initial playing phase, the ERPs were collapsed across reward probabilities into four ERPs based on the expectedness of the reward outcomes, as described below. The search phase yielded two ERPs corresponding to the two token feedback conditions (good casino, bad casino).

Following convention, the reward positivity was measured at channel FCz, where it reaches maximum amplitude (see below), utilizing a difference wave approach that isolated the reward positivity from overlapping ERP components such as the P300 (Holroyd & Krigolson, 2007; Sambrook & Goslin, 2015). The ERPs were averaged across condition as follows: For the initial playing phase for each participant, the ERPs to reward feedback stimuli averaged across games in the bad casino (unexpected reward) were subtracted from the ERPs to no-reward feedback stimuli averaged across games in the good casino (unexpected no-reward), to generate an "unexpected" difference wave (i.e., an unexpected-reward positivity). Likewise, the ERPs to reward feedback stimuli averaged across games in the good casino (expected reward) were subtracted from the ERPs to no-reward feedback stimuli averaged across games in the bad casino (expected no-reward),

to generate an "expected" difference wave (i.e., an expected reward positivity; cf. Holroyd & Krigolson, 2007; Holroyd, Krigolson, Baker, Lee, & Gibson, 2009). Furthermore, the reward positivity elicited by the casino token images in the search phase was likewise measured as a difference wave by subtracting the ERPs elicited by the presentation of the good casino tokens from the ERPs elicited by the presentation of the bad casino tokens. The reward positivity amplitude was then determined by finding the maximum negative deflection in the difference wave from 240 to 340 ms (Sambrook & Goslin, 2015) following feedback onset, separately for the expected and unexpected reward feedback (initial playing phase), which isolated the interaction of expectancy with valence by removing the main effect of probability (Holroyd & Krigolson, 2007; Sambrook & Goslin, 2015), and for the good and bad casino tokens (search phase). The grand-average scalp distribution maps of the reward positivity were adjusted for latency differences across individual participants.

Results

Questionnaires

Following the initial playing phase, all participants correctly indicated on Questionnaire 1 which casino had yielded the larger payoff and reported a stronger preference for the good casino. After completing the second playing phase, all but two participants (23 out of 25) reported on Questionnaire 2 that they had searched for the token associated with the good casino and rated the importance of finding their preferred token as 3.5 out of 5 (±1.1).

Behavior

For the initial playing phase, a repeated measures analysis of variance (ANOVA) on game choice with Casino (good and bad) and Reward Probability (low, medium, and high) as factors revealed a significant main effect of reward probability, $F(2, 48) = 14.1$, $p < .01$, $\eta_p^2 = .37$. Post-hoc tests indicated that participants chose the high-reward games (40% ± 0.1%) more often than the medium-reward games (31% ± 0.1%), $F(1, 24) = 22$, $p < .01$, $\eta_p^2 = .48$, or the low-reward games (29% ± 0.1%), $F(1, 24) = 16.5$, $p = .01$, $\eta_p^2 = .41$; the choices for the latter two games were not significantly different from one another. No other effects were statistically significant. A comparable ANOVA on the response times for the games did not yield statistically significant results.

For the search phase, a paired t test revealed that participants responded faster on trials following the receipt of a good casino token (818 ± 466 ms) than on trials following receipt of a bad casino token (868 ± 440 ms), $t(24) = –2.5$, $p = .02$, Cohen's $d = –0.49$.

Electrophysiology

We first asked whether participants had learned the values of the games within each casino, as reflected in the reward positivity amplitude. Consistent with previous findings (Holroyd & Krigolson, 2007; Holroyd et al., 2009; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; Sambrook & Goslin, 2015), the reward positivity to unexpected outcomes (–6.8 ± 5.6 μV) was significantly larger than the reward positivity to expected outcomes (–3.0 ± 4.3 μV), $t(24) = –3.1$, $p = .01$, Cohen's $d = 0.62$ (Fig. 3a and b) and was distributed over frontocentral areas of the scalp for the unexpected condition (Fig. 3e; Holroyd & Krigolson, 2007; Miltner et al., 1997; Walsh & Anderson, 2011). The reward positivity for the expected condition exhibited a broader distribution over posterior areas (Fig. 3d), which was expected, given its smaller amplitude. These results confirmed that the participants learned the values of the casino games and that the reward positivity reflects an RPE to the game outcomes.

We then asked whether participants had learned the reward values of the casinos themselves, even though the casino contexts were irrelevant to playing the games inside. Relative to the bad casino tokens, the good casino tokens elicited a more positive-going ERP in the time range of the reward positivity (Fig. 3c). The peak amplitude of the difference wave (bad token minus good token ERPs: –3.2 ± 3.1 μV) was statistically different from zero,[3] $t(24) = –5.1$, $p < .01$, Cohen's $d = 1$, and was distributed over frontocentral areas of the scalp (Fig. 3f). These results indicate that the individual rewards received following each game choice during the initial playing phase generalized to the contexts in which the games were played—that is, to the casinos—even though the contexts had no bearing on task performance.

[3] The peak detection algorithm is relatively robust against overlap with other ERP components, as compared to averaging the voltage values within a temporal window (Luck, 2014), and has been commonly used to measure reward positivity amplitude (e.g., Dunning & Hajcak, 2007; Foti & Hajcak, 2009; Hajcak, Moser, Holroyd, & Simons, 2007; Leng & Zhou, 2010; Masaki, Takeuchi, Gehring, Takasawa, & Yamazaki, 2006; Miltner et al., 1997; Onoda, Abe, & Yamaguchi, 2010). Nevertheless, this algorithm provides a biased estimate of the component amplitude, and therefore may overestimate the true effect sizes (Luck, 2014, online chap. 9). For this reason, we confirmed that this statistical effect was real by computing the following "absolute" amplitude measure of the reward positivity difference wave to the token: For each participant, we identified the largest absolute voltage within 240–340 ms, irrespective of the sign of the voltage (positive or negative). If the null hypothesis of no difference between the conditions were true, then across subjects this algorithm would be equally likely to select a positive versus a negative peak, and the value across subjects would not differ statistically from zero. In fact, this analysis resulted in a reward positivity that was significantly more negative than zero ($p = .02$), confirming that the reward positivity was not artifactual.

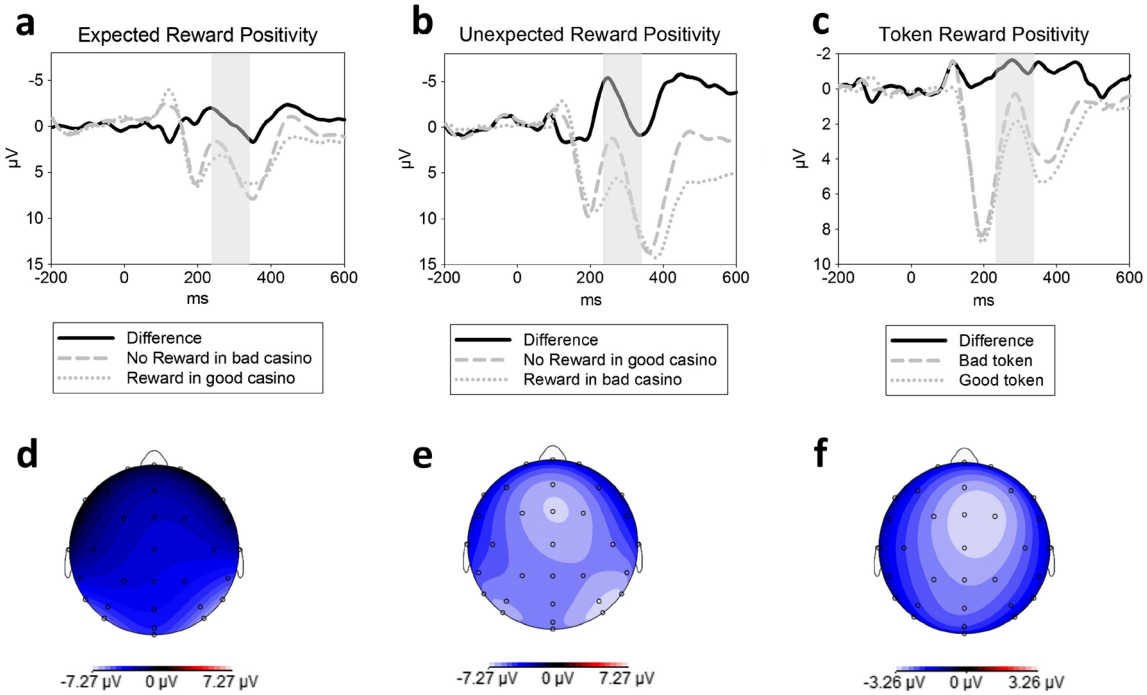

Fig. 3 Event-related brain potentials (ERPs) and the associated scalp voltage maps elicited during the initial playing phase and search phase. (**a**–**c**) Reward positivities measured at FCz in the time window of 240–340 ms, highlighted in gray. Negative is plotted up by convention. During the initial playing phase, (**a**) reward positivity for the expected condition (*thick black line*), elicited by the ERPs to the no-reward feedback in the bad-casino games (*gray dashed line*) and to the reward feedback in the good-casino games (*gray dotted line*), and (**b**) reward positivity for the unexpected condition (*thick black line*), elicited by the ERPs to the no-reward feedback in the good-casino games (*gray dashed line*) and to the reward feedback in the bad-casino games (*gray dotted line*). (**c**) Token reward positivity during the search phase, elicited by the ERPs to the bad-casino tokens (*gray dashed line*) and to the good-casino tokens (*gray dotted line*). (**d**–**f**) Reward positivity scalp distributions. Note that latency differences across the participants were corrected within the 240- to 340-ms time window, so that the peak latency comprising the grand average varied across participants. During the initial playing phase, (**d**) expected reward positivity scalp distribution and (**e**) unexpected reward positivity scalp distribution. (**f**) Token reward positivity scalp distribution during the search phase.

Discussion

An influential theory holds that the reward positivity is produced by the modulatory influence of midbrain dopamine RPE signals on ACC activity (Holroyd & Coles, 2002). Despite accumulating evidence indicating that this ERP component in fact indexes an RPE (Sambrook & Goslin, 2015; Walsh & Anderson, 2012), its purpose in cognitive control remains poorly understood. In particular, the reward positivity amplitude has not consistently been related across studies with trial-by-trial adjustments in behavior, suggesting that the signal does not reflect a simple RL process per se (Holroyd & Umemoto, 2016). Here we asked whether the reward positivity amplitude would be sensitive to the learned reward values of task contexts.

Toward this end, we recorded the ERPs from participants engaged in a novel casino-gambling task in which rewards were obtained more frequently in a "good" casino than in a "bad" casino (see also Diuk et al., 2013). Crucially, whereas the casino games were played repeatedly inside the casinos, the casinos themselves were each selected only once. We found that the reward positivity amplitude was sensitive to differences in reward value associated both with the casino games and with the casinos themselves (Fig. 3), confirming that the reward positivity was associated with the contexts in which those rewards were delivered, even though that contextual information was irrelevant to the optimal policy within each casino (see also Osinsky et al., 2017).

Notably, these data are incompatible with a simple RL mechanism that represents the actions for casino selection and the actions for game selection identically: Computational simulations illustrated that, in such a case, the task parameters that promote learning of the casino values interfere with learning the values of the individual games (supplementary materials and Supplementary Fig. 4). Such an algorithm could solve the problem if the agent were exposed to a series of episodes that paired casino selection with feedback delivery (Botvinick, Niv, & Barto, 2009). However, the task design prevented against that possibility by requiring participants to select each casino only once. Instead, to learn the values of the casinos, participants were required to maintain an internal representation of each casino while playing in it.

For this reason, differential reinforcement of the casinos and the games entails introducing an asymmetry between the two types of actions, in terms of either their representations or their parameter values. Such a difference could be implemented by a variety of mechanisms. For example, a simple RL account that represented each casino as a hidden state (rather than as an action) that reoccurred on each trial could learn to associate the different contexts with reward values. In a recent study by Palminteri, Khamassi, Joffily, and Coricelli (2015), participants learned by trial and error to select between pairs of two stimuli that differed in their expected values. As in the present experiment, the average values of the contexts differed across blocks of trials. Computational simulations revealed that a model incorporating context as a reoccurring hidden state provided a better account of participant behavior than did models without such a state, although it was unclear whether such state values could be used to select one context over the other. Furthermore, the model incorporated different learning rates for the actions and the context, which raises the question of how these parameters were determined.

Although the specific mechanism behind these results remains to be determined, we favor an account based on principles of hierarchical reinforcement learning (HRL). This hypothesis is based in part on evidence indicating that the source of the reward positivity—the ACC—is concerned with regulating task performance in a hierarchical manner, as we have reviewed elsewhere (Holroyd & Yeung, 2012; see also Ribas-Fernandes et al., 2011). Computational simulations that implement these principles account for the effects of ACC damage on the behavior of nonhuman animals; specifically, the values of task contexts are learned by averaging the rewards received during task execution (Holroyd & McClure, 2015). Notably, by separating the values of low-level actions from the values of the task contexts in which they occur, this mechanism naturally prevents the context–action interference described above (see the supplementary materials), while still allowing for high-level action selection consistent with previous reports (Holroyd & McClure, 2015).

Other observations also appear to be inconsistent with an account based on simple RL principles that does not dissociate actions from the contexts in which they occur. As we noted above, the reward positivity amplitude has been inconsistently associated with adjustments in task performance across studies, when consistent results would be expected from a simple RL mechanism (Holroyd & Umemoto, 2016). A simple RL account also runs contrary to the observation that the reward positivity is observed even in the absence of overt, task-related behavior (e.g., Yeung et al., 2005). Furthermore, the reward positivity amplitude decreases with increasing delays between the response and the feedback; even after only a 6-s delay, the difference in the ERPs between wins and losses becomes negligible (Weinberg, Luhmann, Bress, & Hajcak, 2012). In the absence of a contextual signal maintained throughout each casino period, reinforcement at the longest delays—that is, following 90 responses, minutes after the casino was selected—should not have updated the casino values.

Although the reward positivity is believed to be produced by the ACC, it is likely that this process is also supported by other brain areas. In particular, the hippocampus can facilitate the formation of episodic memories of rewarding contexts. The midbrain dopamine system modulates hippocampal activity directly, and reward-related dopamine activity appears to enhance the encoding, retention, and generalization of episodic memories (Shohamy & Adcock, 2010; Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012). In nonhuman-animal studies that pair rewards with the environment, or context, in which they are delivered, the hippocampus has also been implicated in the development of place preferences (Bardo & Bevins, 2000; Tzschentke, 2007). Such considerations suggest that the ACC may contribute to a network of brain areas that support the learning of contextual reward values.

Author note This research was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant and Discovery Accelerator Supplement (312409-05) and a University of Victoria Neuroeducation Network Research Grant awarded to C.B.H.

References

Bardo, M. T., & Bevins, R. A. (2000). Conditioned place preference: What does it add to our preclinical understanding of drug reward? *Psychopharmacology, 153,* 31–43. doi:10.1007/s002130000569

Becker, M. P. I., Nitsch, A. M., Miltner, W. H. R., & Straube, T. (2014). A single-trial estimation of the feedback-related negativity and its relation to BOLD responses in a time-estimation task. *Journal of Neuroscience, 34,* 3005–3012. doi:10.1523/JNEUROSCI.3684-13.2014

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652. doi:10.1037/0033-295X.108.3.624

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition, 113,* 262–280. doi:10.1016/j.cognition.2008.08.011

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684. doi:10.1016/S0022-5371(72)80001-X

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711. doi:10.1038/nn1560

Diuk, C., Tsai, K., Wallis, J., Botvinick, M. M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience, 33,* 5797–5805. doi:10.1523/JNEUROSCI.5445-12.2013

Dunning, J. P., & Hajcak, G. (2007). Error-related negativities elicited by monetary loss and cues that predict loss. *NeuroReport, 18,* 1875–1878. doi:10.1097/WNR.0b013e3282f0d50b

Ebitz, R. B., & Hayden, B. Y. (2016). Dorsal anterior cingulate: A Rorschach test for cognitive neuroscience. *Nature Neuroscience, 19,* 1278–1279. doi:10.1038/nn.4387

Foti, D., & Hajcak, G. (2009). Depression and reduced sensitivity to non-rewards versus rewards: Evidence from event-related potentials. *Biological Psychology, 81,* 1–8. doi:10.1016/j.biopsycho.2008.12.004

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55,* 468–484. doi:10.1016/0013-4694(83)90135-9

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology, 44,* 905–912. doi:10.1111/j.1469-8986.2007.00567.x

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109,* 679–709. doi:10.1037/0033-295X.109.4.679

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

Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective, & Behavioral Neuroscience, 9,* 59–70. doi:10.3758/CABN.9.1.59

Holroyd, C. B., & McClure, S. M. (2015). Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychological Review, 122,* 54–83. doi:10.1037/a0038339

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are refected in the event-related brain potential. *NeuroReport, 14,* 2481–2484. doi:10.1097/01.wnr.0000099601.41403.a5

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,* 688–697. doi:10.1111/j.1469-8986.2008.00668.x

Holroyd, C. B., & Umemoto, A. (2016). The research domain criteria framework: The case for anterior cingulate cortex. *Neuroscience and Biobehavioral Reviews, 71,* 418–443. doi:10.1016/j.neubiorev.2016.09.021

Holroyd, C., & Yeung, N. (2011). An integrative theory of anterior cingulate cortex function: Option selection in hierarchical reinforcement learning. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural basis of motivational and cognitive control* (pp. 333–349). Cambridge, MA: MIT Press.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences, 16,* 122–128. doi:10.1016/j.tics.2011.12.008

Jasper, H. H. (1958). The ten twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology, 10,* 371–375.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., III, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303,* 1023–1026. doi:10.1126/science.1089910

Kolling, N., Wittmann, M. K., Behrens, T. E. J., Boorman, E. D., Mars, R. B., & Rushworth, M. F. S. (2016). Value, search, persistence and model updating in anterior cingulate cortex. *Nature Neuroscience, 19,* 1280–1285. doi:10.1038/nn.4382

Leng, Y., & Zhou, X. (2010). Modulation of the brain activity in outcome evaluation by interpersonal relationship: An ERP study. *Neuropsychologia, 48,* 448–455. doi:10.1016/j.neuropsychologia.2009.10.002

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: MIT Press.

Masaki, H., Takeuchi, S., Gehring, W. J., Takasawa, N., & Yamazaki, K. (2006). Visually induced feeling of touch. *Brain Research, 1105,* 110–121.

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. *Journal of Cognitive Neuroscience, 9,* 788–798. doi:10.1162/jocn.1997.9.6.788

Onoda, K., Abe, S., & Yamaguchi, S. (2010). Feedback-related negativity is correlated with unplanned impulsivity. *Neuroreport, 21,* 736–739. doi:10.1097/WNR.0b013e32833bfd36

Osinsky, R., Ulrich, N., Mussel, P., Feser, L., Gunawardena, A., & Hewig, J. (2017). The feedback-related negativity reflects the combination of instantaneous and long-term values of decision outcomes. *Journal of Cognitive Neuroscience, 29,* 424–434. doi:10.1162/jocn_a_01055

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications, 6,* 8096. doi:10.1038/ncomms9096

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. doi:10.1163/156856897X00366

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology, 52,* 449–459. doi:10.1111/psyp.12370

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron, 71,* 370–379. doi:10.1016/j.neuron.2011.05.042

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin, 141,* 213–235. doi:10.1037/bul0000006

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience, 19,* 1286–1291.

Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences, 14,* 464–472. doi:10.1016/j.tics.2010.08.002

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal–midbrain encoding of overlapping events. *Neuron, 60,* 378–389. doi:10.1016/j.neuron.2008.09.023

Tzschentke, T. M. (2007). Measuring reward with the conditioned place preference (CPP) paradigm: Update of the last decade. *Addiction Biology, 12,* 227–462. doi:10.1111/j.1369-1600.2007.00070.x

Umemoto, A., & Holroyd, C. B. (2016). Exploring individual differences in task switching: Persistence and other personality traits related to anterior cingulate cortex function. *Progress in Brain Research, 229,* 189–212. doi:10.1016/bs.pbr.2016.06.003

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences, 108,* 19048–19053. doi:10.1073/pnas.1117189108

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural

adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews, 36,* 1870–1884. doi:10.1016/j.neubiorev.2012.05.008

Weinberg, A., Luhmann, C. C., Bress, J. N., & Hajcak, G. (2012). Indices of reward processing. *Cognitive, Affective, & Behavioral Neuroscience, 12,* 671–677. doi:10.3758/s13415-012-0104-z

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science, 338,* 270–273. doi:10.1126/science.1223252

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex, 15,* 535–544. doi:10.1093/cercor/bhh153