NeuroImage 183 (2018) 121–131

Contents lists available at ScienceDirect

NeuroImage

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

Electrophysiological responses of medial prefrontal cortex to feedback at different levels of hierarchy

Danesh Shahnazian [a,*], Kurt Shulver [b], Clay B. Holroyd [a]

[a] Department of Psychology, University of Victoria, Canada
[b] Department of Psychology, Macquarie University, Australia

ARTICLE INFO

Keywords:
Medial prefrontal cortex
Reward positivity
Hierarchical reinforcement learning

ABSTRACT

Recent advances in computational reinforcement learning suggest that humans and animals can learn from different types of reinforcers in a hierarchically organised fashion. According to this theoretical framework, while humans learn to coordinate subroutines based on external reinforcers such as food rewards, simple actions within those subroutines are reinforced by an internal reinforcer called a pseudo-reward. Although the neural mechanisms underlying these processes are unknown, recent empirical evidence suggests that the medial prefrontal cortex (MPFC) is involved. To elucidate this issue, we measured a component of the human event-related brain potential, called the reward positivity, that is said to reflect a reward prediction error signal generated in the MPFC. Using a task paradigm involving reinforcers at two levels of hierarchy, we show that reward positivity amplitude is sensitive to the valence of low-level pseudo-rewards but, contrary to our expectation, is not modulated by high-level rewards. Further, reward positivity amplitude to low-level feedback is modulated by the goals of the higher level. These results, which were further replicated in a control experiment, suggest that the MPFC is involved in the processing of rewards at multiple levels of hierarchy.

1. Introduction

Principles of computational reinforcement learning (RL) have successfully accounted for a wide range of behavioral, single cell recordings, lesion and neuroimaging data (Niv, 2009). RL methods are particularly suitable for addressing simple problems characterized by relatively few states and limited numbers of possible actions (Sutton and Barto, 1998). However, many problems faced in real, uncertain environments are much more complex. In particular, humans and other animals typically act in environments associated with a large state-space, which renders many important problems intractable to standard RL algorithms (Botvinick et al., 2009).

Recent developments in RL theory suggest that hierarchical representations can provide a heuristic solution to the problem of scalability (Sutton et al., 1998). Hierarchical representations separate goal-directed behaviors into collections of related sub-goals, each of which is achieved through specific action policies. In this way, the hierarchy of goal and sub-goals corresponds to the hierarchy of tasks and sub-tasks. Note that the question of whether behaviors are represented hierarchically is independent of whether they are also "model-based", which concerns the ability to predict upcoming rewards based on a probabilistic model of the task's state-space (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). The link between model-based and hierarchical behaviors has been extensively discussed elsewhere (Botvinick and Weinstein, 2014; Le Heron, Holroyd, Salamone, & Husain, submitted).

The *options framework* provides a parsimonious solution for hierarchical problems that is relatively similar to existing (flat) RL algorithms. By contrast to flat RL algorithms, which depend only on information about primary rewards, hierarchical reinforcement learning (HRL) entails learning about rewards at multiple levels of abstraction. In the options framework, primary rewards are used to train the system to coordinate sub-tasks in order to achieve a high-level goal (maximizing primary reward), whereas pseudo-rewards are used to select individual actions according to the specific sub-task being executed. For example, given the high-level goal of making a delicious breakfast, the feedback "everything tastes great" to a chef would constitute a primary reward. By contrast, the feedback "you successfully preheated the oven" would constitute a pseudo-reward, as this sub-task is not necessarily rewarding in and of itself.

Option-specific action policies can be learned using standard RL. In these algorithms, unexpected rewards elicit a prediction error signal called a temporal difference error or reward prediction error (RPE)

* Corresponding author.
E-mail address: dshahnaz@uvic.ca (D. Shahnazian).

https://doi.org/10.1016/j.neuroimage.2018.07.064
Received 28 March 2018; Received in revised form 27 July 2018; Accepted 28 July 2018
Available online 4 August 2018

(Sutton and Barto, 1998). In the options framework, the same algorithm utilizes "pseudo-rewards" related to sub-goals in order to generate pseudo-reward prediction errors (pRPEs), which are utilized to optimize action policies that maximize pseudo-reward.

Although the neural mechanisms of RL have been well-studied in recent years (Niv, 2009), the neural correlates of HRL have been relatively less explored. Notably, Botvinick et al. (2009) proposed that the brain regions located in the prefrontal cortex (PFC) enable hierarchical learning. Based on this proposal, orbitofrontal cortex (OFC) and the ventral striatum (VS) evaluate both pseudo-reward and reward information and the dorsolateral striatum (DLS) selects between actions and options accordingly.

To our knowledge, only two studies have addressed this proposal. In a pioneering experiment, Ribas-Fernandes et al. (2011) utilized functional magnetic resonance imaging (fMRI) and the electroencephalogram (EEG) to investigate the neural mechanisms of pseudo-reward learning. In this study, participants used a joystick to play a video game in which they first achieved a sub-task on each trial (by acquiring a package from a specified screen location) and then completed the overall task (by delivering the package to a final destination). Crucially, on some trials the package unexpectedly appeared at a new location during the sub-task phase. This made the sub-goal either harder or easier to achieve without changing the overall difficulty of the primary goal. In this way, the experiment varied the size of pRPEs (elicited by changes in distance to the sub-goal) while simultaneously controlling for the size of the RPEs (elicited by changes in the distance to the goal). The results showed that a region of medial prefrontal cortex (MPFC) called anterior midcingulate cortex (aMCC) was sensitive to pRPEs.

More recently, Diuk et al. (2013) looked at the hemodynamic correlates of learning simultaneously, from different levels of hierarchy, in a computerized casino task. On each trial participants first selected between two casinos (the higher-level choice), and then serially selected two out of four slot machines within that casino (the lower-level choices). Each slot choice was followed by a feedback stimulus indicating the number of points associated with that choice (constituting the pseudo-rewards). Crucially, the second slot-level feedback stimulus was simultaneously shown with another feedback stimulus indicating money won or lost (constituting the primary reward). Subjects were instructed to discover the optimal sequence of choices that would result in maximum monetary gain. The lower-level feedback depended on the sequence of lower-level choices and the higher-level feedback depended on the higher-level choices, which encouraged participants to adapt their behavior according to both levels of feedback. Yet contrary to the study by Ribas-Fernandes et al. (2011), the results failed to reveal pRPEs in the MPFC. Rather, they found only one brain region, the VS, to be involved in both types of learning.

Notably, Holroyd and Yeung (2012) developed a theory of aMCC function that was partly inspired by findings from lesion studies in humans and other animals. They proposed that the aMCC lies at the apex of this hierarchical action selection mechanism. On this view, the MPFC utilizes RPE signals to learn option values, as opposed to the values of the primary actions that comprise the options (Holroyd and McClure, 2015; Holroyd and Umemoto, 2016). Consistent with this idea, Zarr and Brown (2016) found that the MPFC responds to feedback at different levels of hierarchy along its rostro-caudal extent. In this fMRI study, participants learned from feedback whether a rule at a certain level of hierarchy had changed. The results showed that the activity of the rostral parts of the MPFC are modulated by higher-level feedback and the activity of the caudal parts are modulated by lower-level feedback. Although the information conveyed by the feedback used in this study does not map directly onto the feedback used in studies concerning economical decision making, it does point to the involvement of the MPFC in processing different kinds of lower-level information in hierarchically-organized tasks.

While the hierarchical theory of MPFC has helped elucidate the results of several recent empirical studies (Balaguer et al., 2016; Umemoto et al., 2017), strong evidence for this mechanism is lacking. Further, while Holroyd and Yeung (2012) suggest that MPFC learns option values based on (high-level) RPEs, Ribas-Fernandes et al. (2011) observed that MPFC produces (low-level) pRPEs. The role of the MPFC in HRL, and specifically whether or not the MPFC is sensitive to pRPEs and/or RPEs across different task manipulations remains unclear.

To investigate this issue, we recorded EEG from participants engaged in a computerized casino task (see also Krigolson and Holroyd, 2006; Krigolson et al., 2008; Umemoto et al., 2017). The task was inspired by Diuk et al.'s (2013) casino paradigm and evaluated the amplitude of the reward positivity (RewP) to different feedback events. The RewP is a component of the event-related brain potential (ERP) said to reflect RPE signals carried via dopamine projections to the MPFC (Holroyd & Coles, 2002, 2008). A variety of empirical studies have supported this idea. For example, RewP amplitude is sensitive to the probability of rewarding stimuli and is also evoked by stimuli that predict rewarding outcomes (Holroyd and Coles, 2002; for reviews see Holroyd and Umemoto, 2016; Walsh and Anderson, 2012; Sambrook and Goslin, 2015).

In the current study, we asked whether RewP amplitude is sensitive to RPE signals, pRPE signals, or both. Following the experimental paradigm of Diuk et al. (2013), the task featured three consecutive choices (one casino choice followed by two slot choices) and three feedback stimuli (2 outcomes presented as game points followed by 1 monetary outcome) on each trial. The participants were instructed that through trial and error they should find the sequence of choices that would maximize their monetary gain. The task was structured to encourage participants to perceive the first two feedback stimuli on each trial (the slot-level point outcomes) as conveying pseudo-reward information, and the third feedback stimulus (the casino-level monetary outcome) as conveying reward information. Crucially, the reward contingencies were coded such that the first lower-level feedback stimulus on each trial was probabilistically less indicative of higher-level feedback relative to the second lower-level feedback. Moreover, on approximately half of the trials the higher-level feedback was not determined by the lower-level outcomes that preceded it, allowing for the production of (non-zero) high-level prediction errors.

We considered three possible conclusions regarding the overall outcome of the task. First, both low-level feedback and high-level feedback would elicit the RewP. In line with a flat RL account, the RewP would reflect an RPE to primary rewards where the higher-level feedback constitutes the reward. Here, the casino level feedback (wins vs. losses) would elicit a RewP. Further, given that the lower-level feedback is predictive of the higher-level feedback, the RewP should "propagate back in time" during the course of the experiment, as past literature has shown that stimulus cues that predict reward elicit the RewP (Holroyd et al., 2011). Further, because the first slot outcome is less predictive than the second slot outcome of the casino outcome, the second slot outcome should elicit a larger RewP than the first slot outcome (Fig. 1, first row). By contrast, and in line with an HRL account, the RewP could reflect both pRPEs to low-level feedback and RPEs to high-level feedback. In this case, all feedback events would elicit a RewP, and the sizes of RewP to the first and second lower-level feedback stimuli would be comparable (Fig. 1, second row). This possibility differs from the flat RL account in that it predicts a larger RewP to the first slot-machine outcomes.

Second, only the high-level reward, and not the low-level reward, could elicit the RewP (Fig. 1, third row). This possibility is inconsistent with a flat RL account, which predicts that the slot feedback would also come to elicit the RewP (see above). By contrast, this possibility is consistent with an HRL account on the assumption that low-level rewards are processed by a different neural system.

Third, only the low-level reward, and not the high-level reward, would elicit the RewP (Fig. 1, fourth row). This possibility is consistent with both flat RL and HRL accounts. On the flat account, low-level feedback would elicit the RewP if subjects ignored the monetary win/loss feedback and instead viewed the points feedback as intrinsically rewarding, as can occur in video games, for example. Here the slot-level

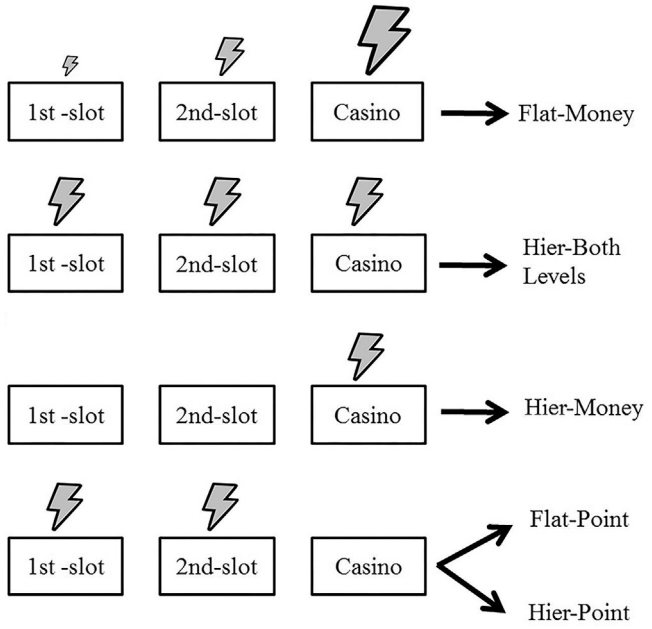

Fig. 1. Possible reward positivity (RewP) results and their compatibility with different hypotheses of hierarchical processing. The lightning symbol represents the RewP to feedback information, with relative size indicating relative RewP amplitude. "Hier-Both Levels" refer to the hypothesis that RewP is sensitive to feedback at both levels of hierarchy. "Hier-Money" refers to the hypothesis that the RewP is only sensitive to high-level feedback. "Flat-Point" and "Hier-Point" refer to the hypotheses that RewP is sensitive to low-level feedback in a flat or hierarchical manner, respectively.

feedback would effectively serve as the primary reward, despite explicit task instructions that frame the task as a hierarchical problem, the goal of which is to win money. Similarly, with a HRL account, the RewP would reflect pRPEs to the first and second slot-machine feedback, which would elicit RewPs of comparable size. However, the casino outcome would fail to elicit a RewP, presumably because the high-level rewards are processed by a different brain area.

We predicted that if subjects frame the problem hierarchically, then their choice behavior should be sensitive to positive vs. negative outcomes at both levels of feedback. Nevertheless, these behavioral adjustments might or might not be associated with RewP amplitude, given observed dissociations between RewP amplitude and learning (Holroyd and Umemoto, 2016). As described below, we first conducted a primary experiment (Experiment 1) to investigate this question, and then conducted a second "control" experiment (Experiment 2) to verify the results.

2. Experiment 1

2.1. Method

2.1.1. Participants

25 participants (20 female; 22 right-handed; aged 18–26, M = 20.8, SD = 1.84) participated in the experiment. All of the participants had normal or corrected-to-normal vision and none reported a history of head injury. Participants were undergraduate students recruited from the University of Victoria. Each received course credit as well as a monetary bonus based on their task performance, as described below. All participants gave informed consent. The study was approved by the local research ethics committee and was conducted in accordance with the ethical standards prescribed in the 1964 Declaration of Helsinki.

2.1.2. Task design

Participants engaged in a computerized task that was a modified version of the casino task (Diuk et al., 2013), coded using Psychophysics Toolbox version 3 for MATLAB. All stimuli were viewed from a distance of about 70 cm (13.9° wide, 9.8° high) and displayed on a 17-inch computer monitor. The visual angle of each stimulus was 6°. The task consisted of four blocks of 38 trials each. On each trial, participants were required to choose between two casinos and subsequently select two out of four slot machines in order to maximize their monetary gain. To be specific, participants selected one of two door images presented on a computer screen, each of which represented an entrance to a virtual "casino", by pressing a corresponding key on a standard QWERTY keyboard (Fig. 2). Participants pressed the "f" key (with the index finger of the left hand) to select the casino on the left side and the "j" key (with the index finger of the right hand) to select the casino on the right side. If no response was issued within 3 s following the onset of the door stimulus, then 1 cent was deducted from their total earnings, the trial was terminated and the next trial began.

One second after selecting a casino, the participants were presented with an image of four different slot machines (distinguishable by colour) positioned near each of the four corners of a rectangle, which represented the environment inside the selected casino. Slot selection involved pressing a character on the keyboard (characters "f", "j", "c", "m") that was spatially consistent with the location of the machine on the screen. Note that participants were instructed to press the "c", "m" with their left and right thumb respectively. Participants were instructed to select two slot machines consecutively. The first time that a participant selected a slot machine, the colour of the other machines turned grey, indicating deactivation, and the appearance of the slot on the selected machine changed to indicate activation (Fig. 2). Then, a number indicating the amount of game points resulting from that selection appeared at the center of the machine. 3 s after the second selection, the trial outcome was indicated by changing the colour of the bar in the middle of screen either to green, indicating a win of 10 cents, or to red, indicating a loss of 10 cents, as appropriate to the choice-reward contingencies described below (Fig. 2)).

If no response was committed at the casino level in the 3 s period after the presentation of the door image, the trial would be discontinued and another trial would begin. In contrast, if no response was committed at the slot machine level for each of the slot machine choices, the trial would discontinue incurring a 1 Canadian cent penalty on the total earnings made by the participants.

Participants were given 5 dollars in credit to spend during the experiment. They were instructed to maximize their monetary gain by choosing the casino "that gives you on average the greatest chance of winning 10 cents" and to find the slot machines "that are most likely to give you 5 points." Thus, the task instructions were intentionally biased to encourage subjects to represent the game hierarchically.

Importantly, the task was designed to dissociate the effects of lower-level reward versus higher-level reward information on the participants' behavior. Each slot machine choice yielded a dichotomous outcome (0 points or 5 points; hereafter, the "bad" and "good" outcomes at the lower-level, respectively) so that on each trial the participant obtained 0, 5 or 10 points across both choices. For both casinos, the payoffs on the four slot machines were associated with different probabilities of yielding 5 points (30%, 40%, 60% and 70%; implemented as pseudorandom variables) with each slot machine being associated with only one of these payoff rates. Thus, the probabilities of winning a low-level reward did not differ across the two casinos.

The casino outcome on each trial depended on the number of points obtained. Feedback indicating 0 points resulted in a certain loss of 10 cents (the "bad" outcome at the higher-level), feedback indicating 10 points resulted in a certain gain of 10 cents (the "good" outcome at the higher-level). Feedback indicating 5 points yielded a probabilistic reward – 40% chance of gaining 10 cents in one casino and a 60% chance of gaining 10 cents in the other casino – resulting in unexpected high-level outcomes on about half of the trials. Given that the probability of winning the lower-level reward did not differ across casinos, subjects had to pay

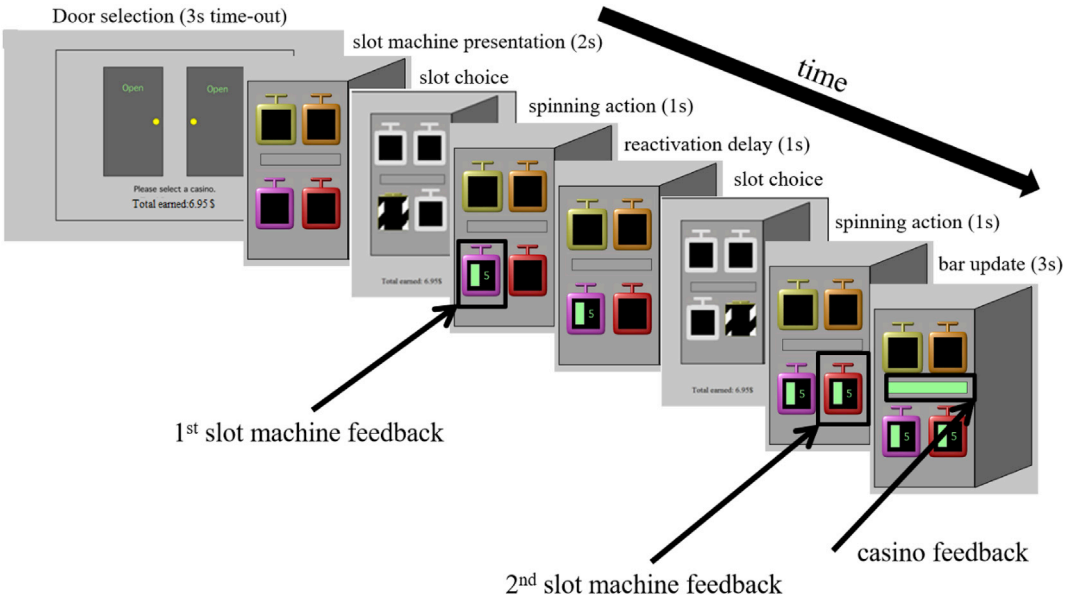

Fig. 2. Event timings for an example trial.

attention to the higher-level reward information in order to maximize their monetary reward (i.e. in which casino they had a higher likelihood of winning money). The identity of the "better" casino was counterbalanced across participants. Note that chance performance on this task results, on average, in 0 cents gain overall.

2.1.3. Behavioral data analysis

Given the explicit hierarchical nature of the task, we hypothesized that the outcomes at each level of hierarchy would only modulate the frequency of choices at that particular level (i.e. casino outcomes modulate casino choices and slot machine outcomes modulate slot machine choices). To test this, we conducted two analyses of variance (ANOVAs) on the frequency of repeating the same action in the future:

To assess whether the casino choices are only modulated by outcomes at the casino level and not by outcomes at the slot machine level, we conducted a two way within-subjects ANOVA with slot machine (bad, good) and casino (bad, good) outcomes as factors on the frequency of repeating the same casino choice (stay probability) on the upcoming trial.

To assess whether the casino choices are only modulated by outcomes at the slot machine level and not by outcomes at the casino level, we conducted a two-way within-subject ANOVA with slot machine (bad, good) and casino outcomes (bad, good) as factors on the frequency of repeating the same slot machine choice (stay probability) on the next trial in which the participant selected the same casino.

2.1.4. Value based computational modeling of choice behavior

In addition, we used a hierarchically organised temporal difference learning (TD) model to account for choice behavior according to a value-based decision making framework. We then compared the model fit for the hierarchical model with that of a TD based flat model. Specifically, we pit two competing models against each other. In one model, choices at both levels of hierarchy were only driven by information at the casino level, which corresponds to a flat agent. In the other model, choices were driven by information from both levels, which corresponds to a hierarchical agent.

2.1.4.1. Flat model. The flat model was implemented using an actor-critic architecture. Here, the critic uses the TD algorithm for evaluating the value of each state while the actor learns the value of each action (Sutton and Barto, 1998). The model only learns from monetary reward feedback; the different points outcomes are represented as different environmental states without intrinsic value.

The agent selects between two, four and three choices on the 1st, 2nd and 3rd steps, respectively. Depending on the choices, the agent can experience either of the two states in step 2 (chosen casino), one of the four states in step 3 (0 vs 5 points feedback in each casino) and one of the eight states in step 4 (delivery of either 0 vs. 5 points feedback following each of the four states on step 2). Hence the agent can experience 2, 4, and 8 states on the 2nd, 3rd and 4th steps of the sequence, respectively.

The value for each state V_F, where V denotes the value and F indicates that this value is assigned by a flat agent, is determined based on the RPE term δ:

$$\boldsymbol{\delta_t} = \boldsymbol{R_{M_{t+1}}} + \boldsymbol{\gamma V_F(S_{t+1})} - \boldsymbol{V_F(S_t)} \tag{1}$$

$$\boldsymbol{V_F(S_t)} \leftarrow \boldsymbol{V_F(S_t)} + \boldsymbol{\alpha\delta_t} \tag{2}$$

Here, S denotes the current state, R_M denotes the monetary reward, γ denotes the discount factor, t denotes the current step of the sequence, and α denotes the learning rate.

The value for each action is also updated based on the prediction error term and actor's learning rate β:

$$\mathbf{h}(\boldsymbol{S_t}, \boldsymbol{a_t}) = \mathbf{h}(\boldsymbol{S_t}, \boldsymbol{a_t}) + \boldsymbol{\beta\delta_t} \tag{3}$$

The probability for selecting each action a is determined according to the softmax equation:

$$\mathrm{P(selecting\ action\ a\ at\ time\ t)} = \frac{\boldsymbol{e}^{\mathbf{h}(\boldsymbol{S_t},\boldsymbol{a_t})/\tau}}{\sum_{\boldsymbol{A_{t,i}}=1}^{\boldsymbol{n}} \boldsymbol{e}^{\mathbf{h}(\boldsymbol{S_t},\boldsymbol{A_{t,i}})/\tau}} \tag{4}$$

In Equation (4), n is the number of possible actions at state S_t and τ is the temperature parameter.

Note that the flat agent does not evaluate slot machine outcomes as having an inherent valence; nevertheless, because each slot machine outcome is represented as a separate state, it can gradually learn that high-point outcomes have higher values, and therefore that actions that lead to high-point outcomes have higher values.

The model depends on four parameters $(\alpha, \beta, \tau, \gamma)$. Because the events in a trial happen in rapid succession, the discount rate is set to 1. The other three parameter values are estimated by fitting the model to each participant's choice behavior using the maximum likelihood estimation procedure (Daw, 2011).

2.1.4.2. Hierarchical model. In this model, door selections are reinforced based on monetary earnings and slot machine selections are reinforced based on points earnings. The model implements an actor-critic learning algorithm separately for each level.

The hierarchical agent learns at two levels of abstraction. These levels are both implemented using an actor-critic architecture. The lower, slot machine level is characterized by four actions and one state; the value for the state V_H, where H indicates that this value is assigned by a hierarchical agent, is learned based on the pRPE term δ_{slot}:

$$\boldsymbol{\delta_{slot}} = \boldsymbol{R_{P_{t+1}}} - \boldsymbol{V_H(S_t)} \tag{5}$$

$$\boldsymbol{V_H(S_t)} \leftarrow \boldsymbol{V_H(S_t)} + \boldsymbol{\alpha_1 \delta_{slot}} \tag{6}$$

In Equation (5), R_P denotes points delivered, i.e., the pseudo-reward, and α_1 is the learning rate for the lower-level critic. The lower-level actor learns the value of selecting each slot machine with β_1 as its learning rate.

$$\mathbf{h}(\boldsymbol{S_t}, \boldsymbol{a_t}) = \mathbf{h}(\boldsymbol{S_t}, \boldsymbol{a_t}) + \boldsymbol{\beta_1 \delta_{slot}} \tag{7}$$

The probability for selecting each action a is determined using the softmax equation with temperature parameter τ_1:

$$\mathrm{P(selecting\ action\ a\ at\ time\ t)} = \frac{\boldsymbol{e}^{\mathbf{h}(S_t, a_t)/\tau_1}}{\sum_{A_{t,i}=1}^{n} \boldsymbol{e}^{\mathbf{h}(S_t A_{t,i})/\tau_1}} \tag{8}$$

In Equation (8), n is the number of possible actions for each state (i.e. $n = 4$ for 1st slot choice, $n = 3$ for 2nd slot choice).

The higher, casino-level is characterized by two actions and one state; the value for the state is learned based on the RPE term.

$$\boldsymbol{\delta_{casino}} = \boldsymbol{R_{M_{t+1}}} - \boldsymbol{V_H(S_t)} \tag{9}$$

$$\boldsymbol{V_H(S_t)} \leftarrow \boldsymbol{V_H(S_t)} + \boldsymbol{\alpha_2 \delta_{casino}} \tag{10}$$

In Equation (9), R_M denotes money delivered, i.e., the primary-reward. α_2 in Equation (10) is the learning rate for the higher-level critic.

The higher-level actor learns the value of each slot machine with β_2 as its learning rate. And the probability for selecting each action is determined using the softmax equation with temperature parameter τ_2:

$$\mathbf{h}(\boldsymbol{S_t}, \boldsymbol{a_t}) = \mathbf{h}(\boldsymbol{S_t}, \boldsymbol{a_t}) + \boldsymbol{\beta_2 \delta_{casino}} \tag{11}$$

$$\mathrm{P(selecting\ action\ a\ at\ time\ t)} = \frac{\boldsymbol{e}^{h(S_t, a_t)/\tau_2}}{\sum_{A_{t,i}=}^{2} \boldsymbol{e}^{h(S_t A_{t,i})/\tau_2}} \tag{12}$$

The model therefore has six parameters: $\alpha_1, \alpha_2, \beta_1, \beta_2, \tau_1, \tau_2$. These parameters are estimated using the maximum likelihood estimation procedure to find the best fit between the model output and the observed participants' choice behavior (Daw et al., 2011).

For model comparison, we used the Akaike information criterion (AIC) (Akaike, 1998) technique to compare the goodness of fit for the two models. The criterion value for this technique includes a penalty that increases with the number of parameters associated with each model. This penalty is intended to lower the chance of overfitting. The criterion value for each model is calculated according to the following equation:

$$\mathrm{AIC\ (model)} = 2\mathrm{k(model)} + 2\ \mathrm{NLL(model)} \tag{13}$$

In Equation (13), k is the number of parameters and *NLL* is the negative log-likelihood (as calculated according to Daw, 2011) of the model. Lower criterion values are associated with better performing models.

2.1.5. EEG acquisition and preprocessing

The EEG signal was acquired using a montage of 33 Ag/AgCl ring electrodes mounted on a nylon electrode cap according to the extended international 10–20 system (Jasper, 1958). The electrode AFz was used as the ground for the EEG recording and the amplifier used an online average reference to amplify the signal. Inter-electrode impedances were maintained below 20 kΩ using an abrasive conductive gel applied to each electrode. The electro-oculogram (EOG) was recorded for the purpose of ocular correction; horizontal EOG was recorded from the external canthi of both eyes, and vertical EOG was recorded from the sub-orbit of the right eye and electrode channel Fp2. Signals were amplified by differential amplifiers with an online bandpass filter in the range frequency response 0.017–67.5 Hz (90 dB per octave roll off) and digitized with a sampling rate of 250 per second. Digitized signals were stored on disk using Brain Vision Recorder software (Brain Products GmbH, Munich).

Post-processing and data visualization were performed using Brain Vision Analyzer software (Brain Products GmbH, Munich). A fourth order digital Butterworth passband filter in the range of 0.1–30 Hz was applied (12 dB per octave roll off for the 0.1 and 24 dB per octave roll off for the 30 Hz). We visually inspected the continuous EEG data for periods of unsystematic artifacts that were unlikely to be corrected by the artifact correction algorithm; any time window containing such data was then rejected from further analysis. On average only 1 percent of the continuous EEG data was rejected based on this procedure (except for one subject who had more than 60% of their data rejected; the entire data set for this participant was then removed from further analysis). Ocular artifacts were corrected using the eye movement correction algorithm described by Gratton et al. (1983). The EEG data were re-referenced to averaged-mastoids electrodes. Epochs of 800 ms duration were created by extracting samples from 200 ms prior to 600 ms following the onset of each feedback stimulus (for the slot machine-level feedback, points shown on the slot machines, and for the casino-level feedback, the green or red colour of the outcome bar) from the continuous EEG, separately for each channel and subject. This procedure created about 450 segments per subject (150 segments for each feedback type (i.e. first lower-level feedback, second lower-level feedback and higher-level feedback) when collapsed across valence condition). Data were baseline-corrected by subtracting the mean voltage during the 200 ms interval preceding feedback stimulus onset in each epoch from the post-feedback stimulus voltages in that epoch. Muscular and other artifacts were removed using a ±100 μV threshold and a ±35 μV step threshold as rejection criteria (Luck, 2014, p.349, 2011). On average 2% of the segments were rejected as a result of the artifact rejection procedure (SD = 2.0%, see supplementary inline material, Appendix B for more details).

To examine the ERPs to lower-level feedback events, six feedback-related ERPs were created for each electrode and subject by averaging the single-trial EEG according to four different outcome types as follows: 1) 5 point and 2) 0 point outcomes to the first slot machine; 3) 5 point and 4) 0 point outcomes to the second slot machine, and 5) 5 point and 6) 0 point outcomes averaged across the first and second slot machine outcomes. RewP amplitude was calculated using a difference-wave approach by subtracting the ERP to the good outcomes from that of bad outcomes for each event of interest, separately for each channel and subject. This approach minimizes overlap with other ERP components (Holroyd and Krigolson, 2007; Sambrook and Goslin, 2015; but see Cavanagh, 2015). RewP at the (lower) slot machine level was calculated separately for 1) the first slot machine outcomes, 2) the second slot machine outcomes, and 3) across both slot machine outcomes.

To examine the ERPs to higher-level outcomes, feedback-related ERPs were created for each electrode and subject by averaging the single-trial EEG separately according to the win outcomes and loss outcomes. Similar to the lower-level, RewP amplitude was calculated by subtracting the ERP to the good outcomes from that of bad outcomes for each event of interest, separately for each channel and subject.

For all inferential statistical analyses, RewP amplitude was evaluated at channel FCz, where previous literature indicates that the component reaches maximum amplitude (Walsh and Anderson, 2012). The frontal-central shape of the distributions was evaluated by inspection. The mean amplitude of the RewP was calculated as the average value of the difference wave observed within the interval 240–340 ms

post-stimulus, as prescribed by a meta-analysis of RewP studies (Sambrook and Goslin, 2015). Hereafter we refer to this interval as the "RewP interval".

We carried out three different analyses to examine which of the three possibilities articulated in the introduction were supported by the evidence. First, a one sample *t*-test on the amplitude of the RewP to the slot machine-level outcomes (averaged across the first and second slot machines) was completed in order to examine whether those outcomes modulated RewP amplitude. Second, we examined whether the casino-level outcomes elicited a RewP by conducting a one sample *t*-test on the amplitude of the RewP to the casino outcomes. Third, we compared the amplitude of the RewP elicited by the first slot machine outcome to that elicited by the second slot machine outcome in order to examine whether temporal proximity of the slot machine outcome to the casino outcome modulated RewP amplitude.

3. Results

3.1. Behavioral data

Fig. 3 (left panel) shows the effect of casino and slot machine outcomes on the mean frequency of repeating (staying with) the same casino choice on the following trial. Average reaction times for the first and second slot machine choices were (M = 738, SD = 224) for the first-choice, (M = 446, SD = 178) ms for the second-choice. Average reaction time for the casino choice was (M = 782, SD = 183) ms. The average reaction time for the first slot machine choice was significantly larger than that of the second choice, $t\ (23) = 9.96$, $p < 0.05$.

A two way within-subjects ANOVA with slot machine (bad, good) and casino (bad, good) outcomes as factors on the frequency of repeating the same casino choice (stay probability) on the upcoming trial revealed a significant main effect of casino outcome, $F\ (1, 23) = 29.6$, $p < 0.05$, but no significant main effect of slot machine outcome, $F\ (1, 23) = 1.72$, $p > 0.05$, and no significant interaction between the two factors, $F\ (1, 23) = 0.04$, $p > 0.05$. Therefore, the casino choices were influenced only by the casino outcomes and not by the slot machine outcomes. Because outcomes at the lower-level of the hierarchy did not effect choices at the higher-level, this result is consistent with a hierarchical learning account.

Fig. 3 (right panel) shows the effect of casino and slot machine outcomes on the mean frequency of repeating the same slot choice upon selecting the same casino (stay probability). A two-way within-subjects ANOVA with slot machine (bad, good) and casino outcomes (bad, good) as factors on the stay probability revealed no significant main effect of casino outcome, $F\ (1, 23) = 1.7$, $p > 0.05$, a significant main effect of slot machine outcomes, $F\ (1, 23) = 34.6$, $p < 0.05$ and no significant interaction between the two factors, $F\ (1, 23) = 2.1$, $p > 0.05$. Therefore, because the low-level choices were influenced only by low-level outcomes and not by high-level outcomes, this result is also consistent with a hierarchical learning account.

3.2. Modeling results

Computational simulations of subjects' choice behavior provide further evidence that the subjects represented the task hierarchically. As mentioned before, each participant selected between 24 different sequences total (2*4*3 choices over the 3 steps). A model based on chance behavior would assign a probability of 1/24 to each possible sequence of choices, which corresponds to an average of 3.178 NLL. If the participant's choices are in fact influenced by the outcomes, then both the flat and hierarchical models predict above-chance behavior, meaning that these models would achieve lower NLL as compared to the average

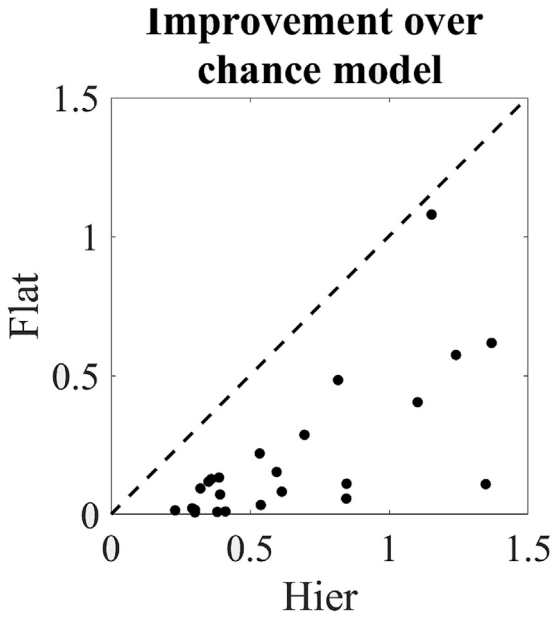

Fig. 4. Comparison between the fit of each model to the choice behavior for Experiment 1. The model fit is calculated by subtracting the negative log-likelihood (NLL) of a model-based on chance from the indicated model's NLL and then dividing the result by the number of trials, for both the hierarchical ("Hier.") and flat models, for each participant.

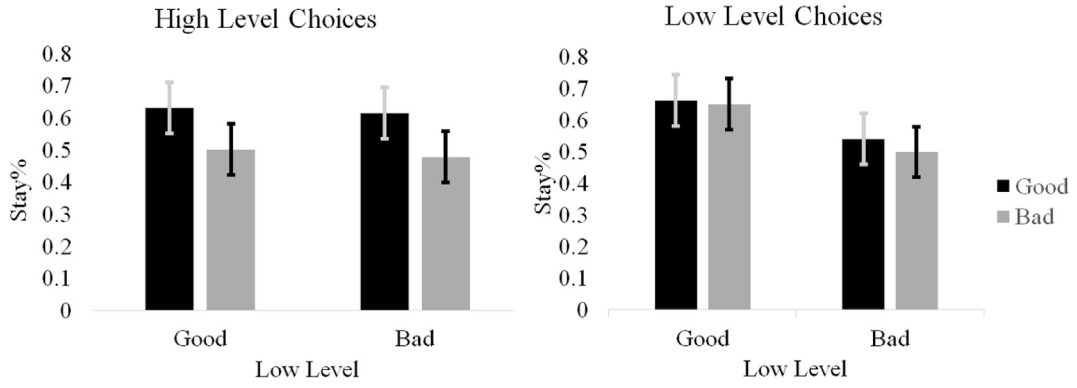

Fig. 3. Feedback at different levels of hierarchy modulate choices at their respective level.
Left panel: probability of repeating (stay%) the high-level (casino) choice on the upcoming trial.
Right panel: probability of repeating (stay%) the low-level (slot machine) choice upon entering the same casino again. Behaviors following presentation of good, high-level (casino) feedback and bad, high-level feedback are depicted in black and grey, respectively. Good and bad low-level (slot machine) feedback are indicated on the x-axis.

performance of a chance model (i.e. a model that selects competing actions with equal probability). Fig. 4 shows the difference in NLLs between the chance model and the flat model, and between the chance model and the hierarchical model, averaged across trials separately for each subject. As revealed by inspection, the hierarchical model fits the data better than the flat model for all participants. Further, AIC criterion values – which control for the number of parameters *k* for each model – revealed superior performance by the hierarchical model, despite the latter having more parameters:

Experiment 1: AIC (flat model)-AIC (hierarchical model) = 888.03

3.3. Electrophysiological data

The data of one participant were removed from all of the analyses due to a large number of trials rejected because of artifact (67%). Fig. 5 (left column) shows the grand average ERPs recorded at channel FCz to the slot-machine level feedback, averaged across the feedback to both slot machine choices, the RewP computed from these ERPs, and the corresponding scalp distribution of the RewP. As expected, the grand average RewP to the slot machine outcomes is positive-going in the interval 240 –340 ms following feedback presentation. A two-tailed one sample *t*-test on the mean area amplitude ($M = -4.23$, $SD = 2.85$) revealed a significant effect, $t(23) = -7.28$, $p < 0.05$, $d = 1.48$. Moreover, visual inspection of the scalp distribution of the difference wave to slot machine outcome confirms that the RewP exhibits a fronto-central distribution (see Fig. 5 top row left column).

Fig. 5 (middle column) shows the grand average ERPs to the casino level feedback at channel FCz, the RewP computed from these ERPs, and the corresponding scalp distribution. Although the grand average RewP is positive-going in the RewP interval, this value was not statistically different from zero ($M = -0.34$, $SD = 2.15$), $t(23) = -0.77$, $p > 0.05$, $d = 0.16$.

Fig. 5 (right column) shows the RewPs elicited to the first slot machine feedback and to the second slot machine feedback and the scalp distribution associated with the difference between these waveforms in the RewP interval. A two-tailed one sample *t*-test on the mean area amplitude in the RewP interval of the difference wave created by subtracting the RewP to the first slot machine outcome from that of second slot machine outcome ($M = -0.45$, $SD = 3.12$) did not reveal any significant effect, $t(23) = -1.93$, $p > 0.05$, $d = 0.14$.

3.4. Exploratory ERP results

We considered whether low-level feedback, high-level feedback, or both would elicit the RewP in a hierarchically-framed reinforcement learning task. Although we found that low-level feedback elicited the RewP, we failed to find a statistically-significant effect of high-level feedback on RewP amplitude. As discussed above, this result is consistent with both flat RL and HRL accounts (Fig. 1, fourth row). To explore these results further, we created RewPs to the second slot machine outcome averaged separately according to the outcomes of the first slot machine.

According to a hierarchical account, if the RewP encodes pRPEs, then each slot machine outcome should be associated with an inherent valence that does not depend on the other outcomes in the trial. Thus, a good outcome following the second slot machine choice should elicit a

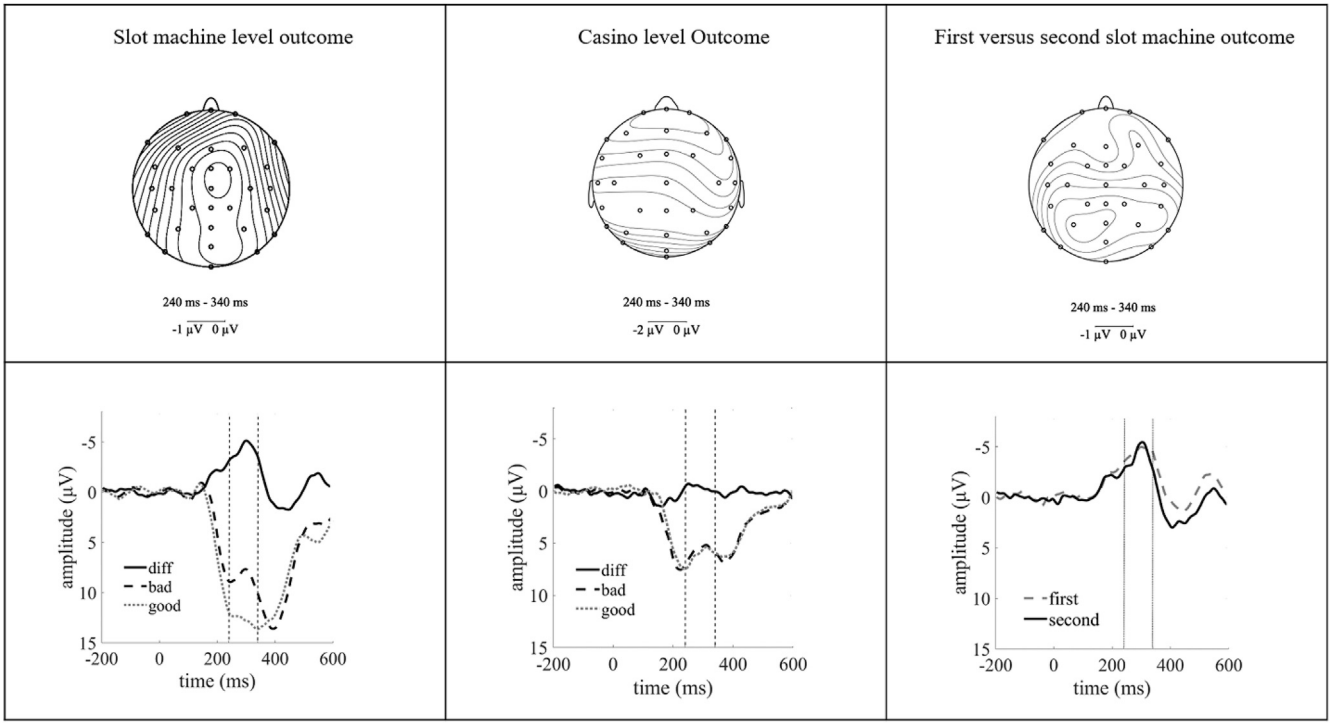

Fig. 5. Reward positivity (RewP) elicited by feedback at different levels of hierarchy.
Left Panels. Bottom: The RewP to lower-level outcomes. Solid line. Difference wave. Dotted line: ERP to the good outcome. Dashed line. ERP to the bad outcome. The vertical dashed lines indicate the RewP interval. Top: Scalp distribution associated with mean RewP amplitude to lower-level outcomes during the RewP interval.
Middle Panels. Bottom: The RewP to higher-level outcomes. Solid line. Difference wave. Dotted line. ERP to the good outcome. Dashed line. ERP to the bad outcome. The vertical dashed lines indicate RewP interval. Top plot. Scalp distribution associated with mean amplitude of the RewP to higher-level outcomes over the RewP interval.
Right Panel. Bottom: RewPs elicited to first and second lower-level outcomes. Solid line. The RewP to the second lower-level outcome. Dashed line. The RewP to the first lower-level outcome. The vertical dashed lines indicate the RewP interval. Top. Scalp distribution associated with mean amplitude of the RewP to higher-level outcomes over the RewP interval.

RewP regardless of the previous slot machine outcome. Fig. 6 (panel a) plots the amplitudes of the raw ERPs to the second slot machine outcome as a function of the first slot machine outcome and Fig. 6 (panel c) plots the associated RewPs. One sample t-tests indicate that when the first slot machine outcome was bad, the second slot machine elicited a RewP ($M = -3.11$, $SD = 3.56$), $t\ (23) = 4.28$, $p < 0.05$, $d = 0.83$. Similarly, when the first slot machine outcome was good, the second slot machine outcome produced a RewP ($M = -6.18$, $SD = 4.74$), $t\ (23) = 6.39$, $p < 0.05$, $d = 1.30$. Importantly, the amplitude of the RewP when the first slot machine outcome was good was significantly higher than when the first slot machine was bad ($M = -3.07$, $SD = 3.90$), $t\ (23) = 3.89$, $p < 0.05$, $d = 0.79$.

These results suggest that good versus bad outcomes to the second slot machine elicit a larger RewP when preceded by good outcomes to the first slot machine as compared to when they are preceded by bad outcomes to the first slot machine.

Further, given the task design, some of the higher-level outcomes were readily predictable from the lower-level outcomes. In particular, subjects could correctly predict whether they would win or lose at the casino level if they had accumulated either 0 or 10 points at the slot-machine level. Because RewP amplitude is enhanced to unpredictable feedback (Sambrook and Goslin, 2015; Walsh and Anderson, 2012), we hypothesized that the predictability of casino outcomes could modulate the amplitude of the RewP. To investigate this possibility, we averaged the mean ERP amplitude to wins versus losses as a function of the total number of points accumulated following the second slot choice, where 0 and 10 points feedback constitute predictable outcomes and 5 points feedback constitutes unpredictable outcomes (Fig. 6, panel b).

A one-sample *t*-test on the amplitude of the RewP to unpredictable casino feedback did not reveal any significant effect ($M = -0.81$, $SD = 3.59$), $t\ (23) = -1.12$, $p > 0.05$, $d = 0.23$. Similarly, a one sample *t*-test on the amplitude of the RewP to predictable casino feedback outcomes did not reveal any significant effect ($M = 0.04$, $SD = 3.32$), $t\ (23) = 0.06$, $p > 0.05$, $d = 0.01$. Further, a *t*-test on the amplitude of the difference between the RewP to unpredictable versus predictable outcomes during the RewP interval did not reveal any effect of predictability on the amplitude of the RewP ($M = -0.86$, $SD = 4.89$), $t\ (23) = 0.86$, $p > 0.05$, $d = 0.17$.

Given the lack of an effect of RewP amplitude to the casino outcomes, we were concerned that subjects might not have paid attention to the

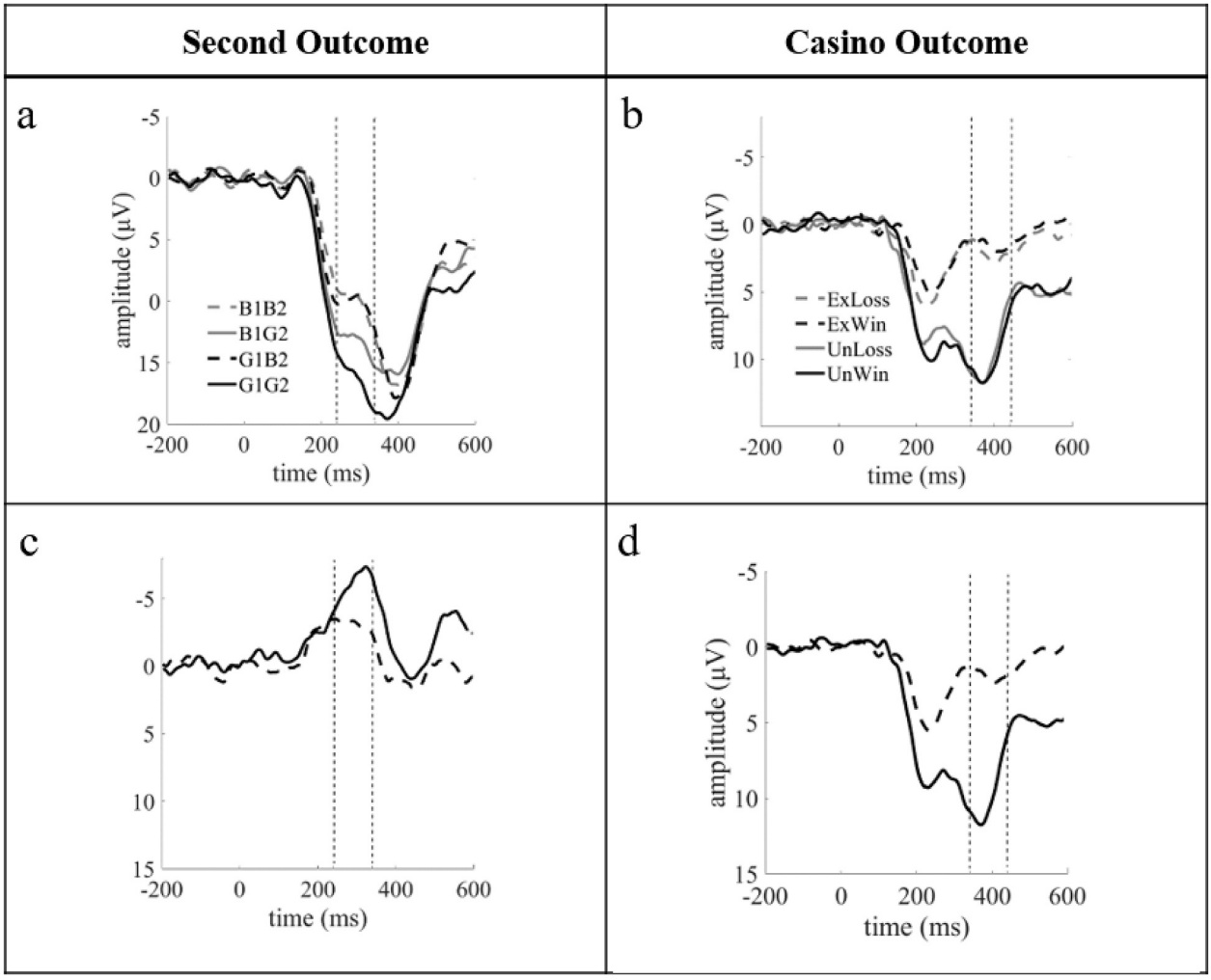

Fig. 6. Post-hoc analyses of the ERPs to the second slot machine outcomes and casino outcomes. a) ERPs to second slot machine outcome as a function of first and second slot machine outcomes. G and B denote Good and Bad outcomes respectively and 1 and 2 indicate feedback following the first and second slot machine choices, respectively. For example, B1G2 indicates a bad outcome following the first slot machine choice and a good outcome following the second slot machine. b) ERPs to the casino outcomes as a function of their valence and expectancy. The prefixes "un" and "ex" denote unexpected and expected outcomes, respectively, and "win" and "loss" indicate wins and loss outcomes, respectively. c) Reward positivity (RewP) elicited by the second slot machine outcome as a function of the first slot machine outcome. Dashed line depicts ERP to a bad outcome following the first slot machine choice and solid line depicts ERP to a good outcome following the first slot machine choice. d) ERPs to unexpected (solid line) vs. expected (dashed line) casino outcomes. The vertical dashed lines indicate the interval of interest (RewP for panels a and c and P3 for panel b and d). Note the different y-axis scales across the panels. All waveforms associated with channel FCz.

casino-level feedback. For this reason, we tested whether the amplitude of the P3 – which is associated with attention and working memory, exhibiting larger amplitudes to infrequent, task-relevant events in a variety of task paradigms (Polich, 2007) – was modulated by the valence and probability of the casino outcomes. We measured the P3 as the mean amplitude of ERP in the interval 340–440 ms post-feedback (which follows the RewP interval immediately and is matched for duration) recorded at channel Pz in accordance with previous literature (Polich, 2007). A two-factor repeated measures ANOVA on P3 amplitude with factors probability (two levels: predictable, unpredictable) and valence (two level: win, loss) revealed a significant main effect of predictability of casino outcome F (1,23) = 77.46, $p < 0.05$,MSE = 56.28. As expected there was no main effect of casino outcome F (1,23) = 1.90, $p > 0.05$, MSE = 26.90, nor an interaction between the two factors F (1,23) = 1.74, $p > 0.05$, MSE = 26.12 (see Fig. 6, panel d for ERPs recorded at channel FCz). This result shows that the P3 is larger to the unexpected casino outcomes (after subjects accumulated 5 points across both slots) than to expected casino outcomes (after subjects accumulated either 0 points or 10 points after both slots), irrespective of whether the outcome constituted a win or a loss.

4. Experiment 2

In the first experiment, we found that low-level feedback elicited the RewP, but we failed to detect a RewP elicited by high-level feedback. On the assumption that high-level feedback in the casino task does not in fact produce a RewP (please see below for a discussion of this inference), then this result is consistent with both flat RL and HRL accounts. According to the flat RL hypothesis, participants might have attributed to the low-level feedback an inherent valence – namely, that a higher number of points is inherently better than a lower number of points – as opposed to reflecting the completion of a sub-goal within the larger goal of maximizing monetary earnings. On this view, people just like to win points; therefore; winning points elicited the RewP. According to the HRL account, subjects represent the task hierarchically, but different brain areas process the different levels of feedback, with only the low-level feedback producing the RewP. To decide between these possibilities, we ran a second experiment in which the participants were told that 0 points resulted in a *win* and 10 points resulted in a *loss* on each trial, and that they should therefore try to minimize the number of points accumulated. A RewP that was more positive-going to 0 points feedback than to 5 points feedback would indicate that the participants construed the feedback in line with the task instructions, as opposed to associating the more desirable feedback with more points. For the purpose of completeness, participants were also told that red bar feedback indicated a win at the casino level (rather than green bar feedback, as was the case for Experiment 1). The experimental methods were otherwise identical to those of Experiment 1. We recorded EEG data from the same sample size (25 participants). Note that similar to Experiment 1, EEG data from one subject were excluded due a high number of trials (>60%) rejected because of excessive EEG artifact.

All of the behavioral and EEG effects reported above for Experiment 1 were replicated in Experiment 2 except for one effect: the second slot machine outcome did not produce a detectable RewP when it followed a bad outcome to the first slot machine choice (Supplementary Inline Materials, Appendix A). The replication of the majority of experimental effects provides evidence for the statistical reliability of the results. Further, the results of Experiment 2 rule out the alternative interpretation that the low-level feedback elicited a RewP simply because subjects inherently like to accumulate points, as in this experiment the RewP was elicited by feedback indicating accumulation of fewer points.

5. Discussion

In the two experiments presented here, we investigated whether an ERP component believed to reflect the involvement of the MPFC in RL, the RewP, is sensitive to reward information at two levels of hierarchy. To this end, we recorded ERPs in two experiments involving choices at two different levels of hierarchy (casino choices and slot machine choices) with explicit feedback pertinent to the level of the task (money) and subtask (points). The behavioral results indicated that participants successfully incorporated information from both levels of feedback into their decisions. Therefore, we infer that subjects represented the task hierarchically.

In contrast, the ERP analysis yielded some surprising results. The most surprising result, which was observed in both experiments, is that although low-level feedback elicited a RewP, the high-level feedback did not elicit a detectable RewP. Furthermore, even though unexpected feedback normally enhances RewP amplitude (Walsh and Anderson, 2012), the RewP to unexpected casino outcomes (i.e., on trials in which the casino outcome could not be inferred from the lower-level outcomes preceding it) also failed to reach statistical significance.

We see three reasons that could give rise to this failure to observe a significant RewP to the casino outcomes. First, the lack of a RewP effect to high-level feedback could simply result from low statistical power. However, the effect sizes associated with RewP to the slot machine outcomes across both experiments were relatively high ($d_z = 1.44$ and 0.50 for Experiments 1 and 2, respectively). With a comparable effect size, our study had relatively high statistical power (around 99% and 61% for Experiments 1 and 2, respectively) to detect whether the RewP is elicited by the casino outcomes.

Second, the task instructions, task design and/or low-level stimulus features could have emphasized the importance of the slot machine outcomes, making them more salient to subjects compared to the casino outcomes. This contrast could have attenuated the size of the RewP to the high-level outcomes, as RewP amplitude is sensitive to the emphasis conveyed by task instructions on different dimensions of feedback information (Nieuwenhuis et al., 2004). As well, the feedback stimuli at the two levels of hierarchy were delivered with different delay periods (with 3 s and 1 s delays for high- and low-level feedback, respectively). Past literature indicates that long delays between the response and subsequent feedback attenuates the size of the RewP (Weinberg et al., 2012; Peterburs et al., 2016; Weismüller and Bellebaum, 2016; Arbel et al., 2017). However, although such elements of the task design could have played a role in attenuating the size of the RewP, the win-stay and hierarchical modeling analyses all indicated that subjects adapted their behavior according to the casino-level outcomes. Moreover, P3 amplitude was elevated to the unexpected casino outcomes relative to the expected casino outcomes. These results indicate that participants paid attention to the casino outcomes and incorporated that information into their decisions. In short, the casino outcomes modulated the participants' choice behavior despite failing to modulate RewP amplitude in a detectable manner. This finding is consistent with substantial evidence indicating a dissociation between RewP amplitude and adaptive behavior (Holroyd and Umemoto, 2016).

A third, not mutually exclusive possibility is that in hierarchical tasks the RewP responds only to lower-level rewards related to subtasks. Given that the first two possibilities are implausible, we believe that the data support this third possibility.

The finding that RewP amplitude is sensitive mainly to lower-level feedback can support two competing accounts about the underlying computational process. First, according to an HRL account, the RewP is sensitive to lower-level feedback events because these events elicit pRPEs. Second, according to a flat RL account, subjects in fact, treated the low-level rewards as high-level (primary) rewards, i.e., they evaluated points as being intrinsically rewarding. However, the results of Experiment 2 ruled out this possibility. In this experiment, the stimuli for good versus bad feedback were swapped, such that 0 points feedback predicted a win at the casino level with higher probability as compared to 5 points feedback. Despite this task manipulation, the RewP continued to be elicited by good (0 point) outcomes. These results indicate that the RewP's sensitivity to the lower-level feedback stimuli was determined by the task

instructions, not by subject preferences or bias. Overall, these results provide evidence that the MPFC is sensitive to lower-level outcomes of hierarchically represented subtasks.

Past research supports this interpretation. Ribas Fernandez et al. (2011) showed that the RewP is sensitive to negative pseudo-reward information in a task that modulated participant effort levels to complete each trial. Their study also involved an fMRI experiment that implicated the MPFC as one of the brain areas responsive to pseudo-reward information.

Comparably, Zarr and Brown (2016) found that the MPFC learns from feedback at different levels of hierarchy along a rostro-caudal gradient, which points to the involvement of the MPFC in processing different types of lower-level information associated with hierarchically-organized tasks. Further, other theoretical treatments propose that different regions of MPFC process feedback at different levels of hierarchy (Holroyd and McClure, 2015). It is possible that whereas neurons in caudal MPFC are oriented in a way that is conducive to generating the RewP as recorded at the scalp (Holroyd and Coles, 2002), neurons in rostral MPFC are less so.

However, our exploratory analyses on the ERPs to lower-level outcomes in Experiment 1, which were replicated in Experiment 2, complicate these conclusions. Notably, the RewP to the second slot machine outcome was modulated by the preceding outcome: For both experiments, the ERP to good feedback to the second slot-machine choice was significantly more positive-going when the event was preceded by good feedback to the first slot-machine choice than by bad feedback to the first slot machine choice. This result is inconsistent with a strict hierarchical account in which each slot machine outcome constitutes the end of an independent episode, in which case the RewP to the second feedback stimulus would not depend on the value of the preceding feedback stimulus. Rather, it appears that the best possible outcome at the lower-level elicited a significantly larger RewP.

These findings are reminiscent of the results of Osinsky et al. (2017), who reported a similar interaction between different levels of task goals in modulating RewP amplitude. In this study, researchers examined the RewP in a task in which outcomes indicating monetary wins versus losses sometimes accorded with (termed task-supportive) or conflicted with (termed task-unsupportive) the long-term task goals. The investigators found that only the task-supportive wins elicited the RewP, indicating sensitivity to both the immediate and long-term action-outcome contingencies. In light of this study, one can speculate that the amplitude of the RewP is enhanced when different levels of goals are achieved simultaneously, as in our task, in which the 10 points outcome signaled achieving goals at two levels of hierarchy.

We should note that, given the inverse problem, there is some controversy about where the RewP is produced. Nevertheless, there is good reason to believe that MPFC – and likely specifically aMCC – is the source. First, several source localization studies converge on a source in the MPFC as the neural generator giving rise to the difference ERPs to positive and negative feedback (see Walsh and Anderson, 2012 for review). According to Walsh and Anderson (2012) only small fraction of the studies have localized the RewP to sources outside of MPFC. Moreover, converging evidence from simultaneous EEG/fMRI recording, a transcranial direct current stimulation experiment and intracranial recordings in rats also point to aMCC as the likely neural generator of the RewP (see Holroyd and Umemoto, 2016 for review). Therefore, while the assertion that the MPFC is the neural generator of RewP is still an open question, it remains the most parsimonious hypothesis.

Overall, the results of our experiments indicate that ERPs believed to be generated in the MPFC are sensitive to pseudo-reward information. Given that the RewP is said to be generated by phasic dopamine release in the MPFC (Holroyd and Coles, 2002), these results resonate with the speculations of Botvinick et al. (2009) that dopamine is involved in signaling pRPEs. Moreover, given that the RewP is only elicited to the second slot machine feedback when that feedback signals a sure win, our results support the possibility that different levels of task goals interact in modulating RewP amplitude.

Acknowledgement

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, [Discovery Grant RGPIN 312409-05].

Cette recherche a été financée par le Conseil de Recherches en Sciences Naturelles et en Génie du Canada (CRSNG) [Discovery Grant RGPIN 312409-05].

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.07.064.

References

Akaike, H., 1998. Information Theory and an Extension of the Maximum Likelihood Principle. In Selected Papers of Hirotugu Akaike. Springer, New York, NY, pp. 199–213.

Arbel, Y., Hong, L., Baker, T.E., Holroyd, C.B., 2017. It's all about timing: an electrophysiological examination of feedback-based learning with immediate and delayed feedback. Neuropsychologia 99, 179–186.

Balaguer, J., Spiers, H., Hassabis, D., Summerfield, C., 2016. Neural mechanisms of hierarchical planning in a virtual subway network. Neuron 90 (4), 893–903. https://doi.org/10.1016/j.neuron.2016.03.037.

Botvinick, M., Weinstein, A., 2014. Model-based hierarchical reinforcement learning and human action control. Phil. Trans. R. Soc. B 369 (1655), 20130480.

Botvinick, M.M., Niv, Y., Barto, A.C., 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113 (3), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011.

Cavanagh, J.F., 2015. Cortical delta activity reflects reward prediction error and related behavioral adjustments, but at different times. Neuroimage 110, 205–216.

Daw, N.D., 2011. Trial-by-trial data analysis using computational models. Decision making, affect, and learning: Atten. Perform. XXIII 23, 3–38.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69 (6), 1204–1215.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., Niv, Y., 2013. Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. J. Neurosci. 33 (13), 5797–5805.

Gratton, G., Coles, M.G.H., Donchin, E., 1983. A new method for off-line removal of ocular artifact. Electroencephalogr. Clin. Neurophysiol. 55 (4), 468–484. https://doi.org/10.1016/0013-4694(83)90135-9.

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

Holroyd, C.B., Coles, M.G., 2008. Dorsal Medial Prefrontal Cortex integrates reinforcement history to guide voluntary behavior. Cortex 44 (5), 548–559.

Holroyd, C.B., Krigolson, O.E., 2007. Reward prediction error signals associated with a modified time estimation task. Psychophysiology 44 (6), 913–917.

Holroyd, C.B., McClure, S.M., 2015. Hierarchical control over effortful behavior by rodent medial frontal cortex: a computational model. Psychol. Rev. 122 (1), 54.

Holroyd, C.B., Umemoto, A., 2016. The research domain criteria framework: the case for Medial Prefrontal Cortex. Neurosci. Biobehav. Rev. 71, 418–443. https://doi.org/10.1016/j.neubiorev.2016.09.021.

Holroyd, C.B., Yeung, N., 2012. Motivation of extended behaviors by medial prefrontal cortex. Trends Cognit. Sci. 16 (2), 122–128.

Holroyd, C.B., Krigolson, O.E., Lee, S., 2011. Reward positivity elicited by predictive cues. Neuroreport 22 (5), 249–252.

Jasper, H.H., 1958. The ten twenty electrode system of the international federation. Electroencephalogr. Clin. Neurophysiol. Suppl. 10, 371–375.

Krigolson, O.E., Holroyd, C.B., 2006. Evidence for hierarchical error processing in the human brain. Neuroscience 137 (1), 13–17.

Krigolson, O.E., Holroyd, C.B., Van Gyn, G., Heath, M., 2008. Electroencephalographic correlates of target and outcome errors. Exp. Brain Res. 190 (4), 401–411.

Le Heron, C., Holroyd, C. B., Salamone, J., D. Husain, M. (U). Brain Mechanisms Underlying Apathy.

Nieuwenhuis, S., Yeung, N., Holroyd, C.B., Schurger, A., Cohen, J.D., 2004. Sensitivity of electrophysiological activity from medial frontal cortex to utilitarian and performance feedback. Cerebr. Cortex 14 (7), 741–747.

Niv, Y., 2009. Reinforcement learning in the brain. J. Math. Psychol. 53 (3), 139–154.

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. J. Cognit. Neurosci.

Peterburs, J., Kobza, S., Bellebaum, C., 2016. Feedback delay gradually affects amplitude and valence specificity of the feedback-related negativity (FRN). Psychophysiology 53 (2), 209–215.

Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118 (10), 2128–2148.

Ribas-Fernandes, J.J., 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 (2), 370–379.

Sambrook, T.D., Goslin, J., 2015. A Neural Reward Prediction Error Revealed by a Meta-analysis of ERPs Using Great Grand Averages. American Psychological Association. Retrieved from: http://psycnet.apa.org/journals/bul/141/1/213/.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: an Introduction, vol. 1. MIT press Cambridge. Retrieved from: http://www.cell.com/trends/cognitive-sciences/pdf/S1364-6613(99)01331-5.pdf.

Sutton, R.S., Precup, D., Singh, S.P., 1998. Intra-option Learning about Temporally Abstract Actions. In ICML, vol. 98, pp. 556–564. Retrieved from: https://www.researchgate.net/profile/Doina_Precup/publication/221344978_Intra-Option_Learning_about_Temporally_Abstract_Actions/links/0912f506316e3065ab000000.pdf.

Umemoto, A., HajiHosseini, A., Yates, M.E., Holroyd, C.B., 2017. Reward-based contextual learning supported by medial prefrontal cortex. Cognit. Affect Behav. Neurosci. 17 (3), 642–651. https://doi.org/10.3758/s13415-017-0502-3.

Walsh, M.M., Anderson, J.R., 2012. Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. Neurosci. Biobehav. Rev. 36 (8), 1870–1884.

Weinberg, A., Luhmann, C.C., Bress, J.N., Hajcak, G., 2012. Better late than never? The effect of feedback delay on ERP indices of reward processing. Cognit. Affect Behav. Neurosci. 12 (4), 671–677.

Weismüller, B., Bellebaum, C., 2016. Expectancy affects the feedback-related negativity (FRN) for delayed feedback in probabilistic learning. Psychophysiology 53 (11), 1739–1750.

Zarr, N., Brown, J.W., 2016. Hierarchical error representation in medial prefrontal cortex. Neuroimage 124, 238–247.