Brain Research 1719 (2019) 183–193

Contents lists available at ScienceDirect

Brain Research

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

Research report

Ready, set, explore! Event-related potentials reveal the time-course of exploratory decisions

Cameron D. Hassall[a], Craig G. McDonald[b], Olave E. Krigolson[a,*]

[a] Centre for Biomedical Research, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada
[b] Department of Psychology, George Mason University, Fairfax, VA 22030, USA

HIGHLIGHTS

- Decisions to explore are preceded by an enhanced feedback-locked P300.
- The reward positivity does not distinguish explorations from exploitations.
- Explorations involve more response conflict than exploitations.

ARTICLE INFO

Keywords:
Explore/exploit dilemma
Reward positivity
N200
P300
Response conflict

ABSTRACT

The decision trade-off between exploiting the known and exploring the unknown has been studied using a variety of approaches and techniques. Surprisingly, electroencephalography (EEG) has been underused in this area of study, even though its high temporal resolution has the potential to reveal the time-course of exploratory decisions. We addressed this issue by recording EEG data while participants tried to win as many points as possible in a two-choice gambling task called a two-armed bandit. After using a computational model to classify responses as either exploitations or explorations, we examined event-related potentials locked to two events preceding decisions to exploit/explore: the arrival of feedback, and the subsequent appearance of the next trial's choice stimuli. In particular, we examined the feedback-locked P300 component, thought to index a phasic release of norepinephrine (a neural interrupt signal), and the reward positivity, thought to index a phasic release of dopamine (a neural prediction error signal). We observed an exploration-dependent enhancement of the P300 only, suggesting a critical role of norepinephrine (but not dopamine) in triggering decisions to explore. Similarly, we examined the N200/P300 components evoked by the appearance of the choice stimuli. In this case, exploration was characterized by an enhancement of the N200, but not P300, a result we attribute to increased response conflict. These results demonstrate the usefulness of combining computational and EEG methodologies, and suggest that exploratory decisions are preceded by two characterizing events: a feedback-locked neural interrupt signal (enhanced P300), and a choice-locked increase in response conflict (enhanced N200).

1. Introduction

Making choices involves managing a trade-off between different decision types, such as risky versus safe, emotional versus logical, and automatic versus deliberative. One such trade-off is deciding whether to exploit previous learning or explore new options (the "explore-exploit dilemma": Gittins and Jones, 1974). Exploration is useful when it reduces our uncertainty about the world and leads to better future outcomes (Behrens et al., 2007). However, in order to experience those positive outcomes, it is also important to exploit what is known, i.e., to forgo exploration in order to make value-maximizing decisions. Humans, like other animals, have evolved neural systems to manage the explore/exploit dilemma, a critical ability in uncertain environments.

Broadly speaking, two neurotransmitters are thought to regulate the explore/exploit dilemma: dopamine and norepinephrine. There is evidence that greater tonic dopamine is associated with exploration (Beeler, 2012; Beeler et al., 2010; Frank et al., 2009; Kayser et al., 2015). For example, individuals with greater dopamine levels in prefrontal cortex tend to explore more (Frank et al., 2009). In addition to dopamine, the neurotransmitter norepinephrine has been implicated in exploration (Aston-Jones and Cohen, 2005; Gilzenrat et al., 2010; Jepma and Nieuwenhuis, 2011; Kane et al., 2017; Warren et al., 2017).

[*] Corresponding author at: Centre for Biomedical Research, University of Victoria, P.O. Box 1700 STN CSC, Victoria, British Columbia V8W 2Y2, Canada.
E-mail address: krigolson@uvic.ca (O.E. Krigolson).

https://doi.org/10.1016/j.brainres.2019.05.039
Received 16 February 2019; Received in revised form 7 May 2019; Accepted 28 May 2019
Available online 29 May 2019

Neurons within the locus coeruleus (LC), the main source of norepinephrine in the brain, show two patterns of firing: phasic bursts of activation in response to task-relevant events, and more gradual tonic (baseline) changes. For example, during a reversal learning task phasic LC activation to a previous target decreases when that target is no longer rewarding; activation shifts instead to the new target (Aston-Jones et al., 1997). Thus, phasic LC activation is associated with good signal detection and stimulus-response learning in monkeys (Aston-Jones et al., 1997; Clayton et al., 2004). An increase in tonic LC activation, on the other hand, is associated with poor task performance and high levels of distraction (Aston-Jones and Cohen, 2005). The tonic LC mode may not be maladaptive, however. Converging animal, drug, and pupillometry evidence suggests that high tonic norepinephrine may promote exploration: trying other bandits in a multi-armed bandit task (Jepma and Nieuwenhuis, 2011), leaving a patch while foraging (Kane et al., 2017), and disengaging from a tone discrimination task when rewards diminish (Gilzenrat et al., 2010).

Investigations into the role of dopamine and norepinephrine in the explore/exploit dilemma have thus far been fruitful. It is therefore surprising that little is known about the electroencephalographic (EEG) correlates of these decisions. This is surprising for two reasons. First, the high temporal resolution of EEG lends itself to the time-course of human decision-making (Heekeren et al., 2008). Second, there is evidence that the activity of dopamine and norepinephrine may be indirectly measured via event-related potentials (ERPs) – the averaged EEG response to an event. For example, the reward positivity is an ERP component thought to reflect the effect of phasic dopamine on anterior cingulate cortex (ACC: Holroyd and Coles, 2002; Holroyd and Yeung, 2012). According to Holroyd and Coles (2002), phasic changes in dopamine signify reinforcement learning (RL) prediction errors that modulate the magnitude of the reward positivity. The ACC, according to this view, is attempting to learn the value of options (sequences of actions: Holroyd and McClure, 2015; Holroyd and Yeung, 2012). Note that the reward positivity is usually thought of as being sensitive to phasic, not tonic, dopamine activity. There is evidence, however, that these two types of dopamine activity are related (Grace et al., 2007; Niv et al., 2007). Relevant here, the reward positivity is affected by tonic dopamine; greater prefrontal baseline dopamine activity predicts either a decreased reward positivity (Marco-Pallarés et al., 2009) or an increased reward positivity (Foti and Hajcak, 2012).

The reward positivity is actually a special case of another ERP component, the N200 (Baker and Holroyd, 2011; Holroyd et al., 2008). While the reward positivity occurs specifically in response to feedback, the N200 is elicited by any task-relevant event, is enhanced for surprising events, and is thought to reflect cortical activity arising from a phasic release of norepinephrine (Hong et al., 2014; Mückschel et al., 2017; Warren and Holroyd, 2012; Warren et al., 2011). Thus, assuming that feedback is unexpected, the amplitude of the reward positivity depends on both reward-related phasic dopamine activity and surprise-related norepinephrine activity. N200 modulation, on the other hand, is tied more to norepinephrine activity alone (Hong et al., 2014; Mückschel et al., 2017; Warren and Holroyd, 2012; Warren et al., 2011). The N200 is often followed by another norepinephrine-dependent ERP component called the P300 (Nieuwenhuis et al., 2005). Like the N200, the P300 is enhanced for infrequent and/or task-relevant stimuli and has also been linked to the phasic release of norepinephrine (Murphy et al., 2011; Nieuwenhuis et al., 2005; Nieuwenhuis et al., 2011). In summary, it may be possible to track phasic changes in norepinephrine via the N200 and P300, and phasic changes in dopamine via the reward positivity.

Previous work on the EEG correlates of exploration and exploitation is sparse. Early work by Bourdaud et al. (2008) analyzed EEG recorded from participants performing a four-armed bandit task (Daw et al., 2006). Bouraud and colleagues (2008) asked simply whether or not pre-response EEG was capable of differentiating decisions to explore and exploit. To answer this question, they showed that machine learning could successfully classify trials as explorations and exploitations based on the frequency content of EEG at frontal and parietal sites (also see Tzovara et al., 2012). Consistent with this result, Cavanagh et al. (2011) observed a correlation between uncertainty and response-locked medial frontal theta power that was positive for exploratory decisions, but negative for exploitative decisions. Finally, Hassall et al. (2013) observed an enhancement of the P300 component at the time of exploratory responses compared to exploitative responses during a sequential risk-taking task called the Balloon Analogue Risk Task (BART: Lejuez et al., 2002). Responses and feedback occur simultaneously in the BART, though, so it is unclear which event (response or feedback) led to the P300 effect observed by Hassall and colleagues (2013).

Our goal here was to use EEG to affirm the roles of dopamine and norepinephrine in managing the explore/exploit dilemma. To do this, we examined ERP components locked to two events in a two-armed bandit task: the (feedback-locked) reward positivity/P300 and the (choice-locked) N200/P300. We hypothesized that the enhanced tonic dopamine activity associated with exploration would, when combined with the usual reward-related phasic dopamine activity, effect the reward positivity (either enhance it or reduce it). In light of conflicting reports (Foti and Hajcak, 2012; Marco-Pallarés et al., 2009) we did not hypothesize as to which decision type would elicit the larger reward positivity, only that there would be a difference. To generate our N200/P300 hypothesis, we considered two somewhat conflicting viewpoints on the role of norepinephrine in regulating the explore-exploit dilemma. As mentioned, previous studies have suggested that the tonic mode of LC activity (low task performance/high distraction) may facilitate exploration, while the phasic mode of LC activity (high task performance/low distraction) may facilitation exploitation (Gilzenrat et al., 2010; Jepma and Nieuwenhuis, 2011; Kane et al., 2017; Nieuwenhuis et al., 2005; Warren et al., 2017). Based on this interpretation, and since the N200/P300 complex is associated with phasic norepinephrine, one might predict enhancements of those components around the time of exploitations. However, Dayan and Yu (2006) interpreted phasic norepinephrine as a neural interrupt signal, signaling a need to update one's model of the world – or context – and to switch strategies accordingly (also see: Bouret and Sara, 2005; Yu and Dayan, 2005). Indeed, Donchin's (1981) context-updating hypothesis of the P300 can be considered a precursor to the LC-NE hypothesis of the P300 (the LC-NE P3 theory: Nieuwenhuis et al., 2005) as both highlight the motivational/task significance of a stimulus. Under this view, a phasic burst of NE (and large, concomitant P300) to feedback could signal a need to explore the environment, provided that exploitation had been the dominant strategy, such as in a reversal learning task in which reversals are relatively rare (as seen in Aston-Jones et al., 1997, for example). For example, in the BART, used by Hassall et al. (2013), explorations were rare, and associated with a greater P300 compared to exploitations. For these reasons, we hypothesized that explorations in our task would be associated with an enhancement of the feedback-locked P300 and choice-locked N200/P300.

2. Results

2.1. Model

Our greedy model generated an average negative log-likelihood of 787, 95% CI [602, 973]. Using softmax for action selection resulted in an improved model fit - a negative log-likelihood of 368, 95% CI [323, 413], $t(17) = -5.94$, $p < .001$, Cohen's d = -1.40. The average tuned softmax model parameters were as follows: $\tau = 0.07$, 95% CI $[-0.04$ $0.17]$ and $\alpha = 0.14$, 95% CI $[-0.02, 0.30]$.

2.2. Behavioural

The mean accuracy (all trials) was 78%, 95% CI [75, 82]. The mean proportion of explorations (all trials) was 20%, 95% CI [17, 24]. Mean

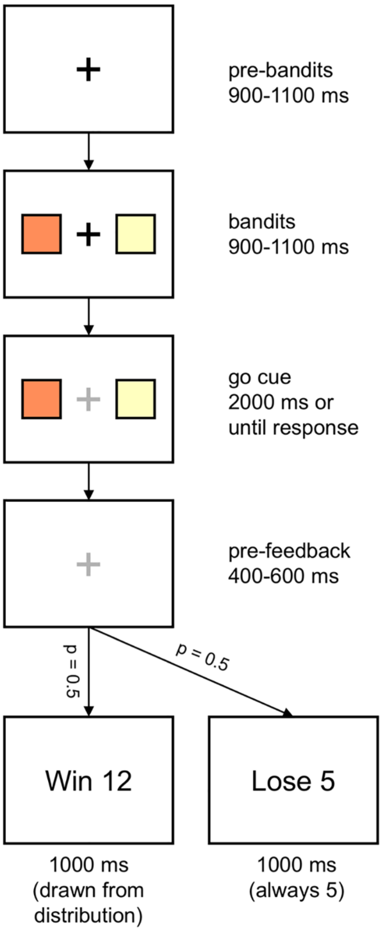

Fig. 1. Two-armed bandit task. Participants were given feedback after selecting one of two colored squares, or bandits. On average, one bandit paid more points than the other. Losses were always the same magnitude (5 points).

accuracy was correlated with mean proportion of explorations, $r(16) = -0.79$, $p < .001$. Explorations and exploitations did not differ in response time exploitations: 376 ms, 95% CI [327, 424], explorations: 371 ms, 95% CI [324, 418]), $t(17) = 1.44$, $p = .16$, Cohen's $d = 0.29$. See Fig. 2 for behavioural results.

2.3. Reward positivity

Single-sample t-tests revealed reward positivities prior to decisions to exploit (3.25 μV, 95% CI [1.89 4.60], $t(17) = 9.74$, $p < .001$, Cohen's $d = 2.30$) and decisions to explore (3.27 μV, 95% CI [2.56 3.98], $t(17) = 5.04$, $p < .001$, Cohen's $d = 1.19$. A paired-samples *t*-test comparing the pre-explore reward positivity to the pre-exploit reward positivity revealed no effect of decision type: $t(17) = -0.05$, $p = .96$, Cohen's $d = -0.01$ (Fig. 3).

2.4. Feedback-Locked P300

Mean P300 was greater for wins than losses prior to both decisions to exploit (pre-exploit loss: 6.76 μV, 95% CI [4.99, 8.54]; pre-exploit win: 10.90 μV, 95% CI [8.91, 12.90]) and decisions to explore (pre-explore loss: 8.55 μV, 95% CI [6.48, 10.61]; pre-explore win: 12.51 μV, 95% CI [10.20, 14.81]). A 2X2 ANOVA with feedback (loss, win) and decision (exploit, explore) as repeated measures revealed main effects of feedback, $F(1,18) = 47.28, p < .001$, $\eta_p^2 = 0.736$, $\eta_g^2 = 0.205$, and, importantly, decision type, $F(1,17) = 18.43$, $p < .001$, $\eta_p^2 = 0.520$, $\eta_g^2 = 0.043$. There was also no significant interaction between feedback and decision, $F(1,17) = 0.14$, $p = .86$, $\eta_p^2 = 0.008$, $\eta_g^2 < 0.001$ (Figs. 4 and 6).

2.5. Single-Trial analysis

The mean slope of a regression line relating P300 magnitude to softmax probability was −0.059, 95% CI [−0.076 to 0.042]. In other words, the P300 dropped by 0.059 μV for every percent of softmax probability. A single-sample *t*-test showed the slope to be significantly different from zero, $t(17) = -7.31, p < .001$, Cohen's $d = -1.72$ (Fig. A1).

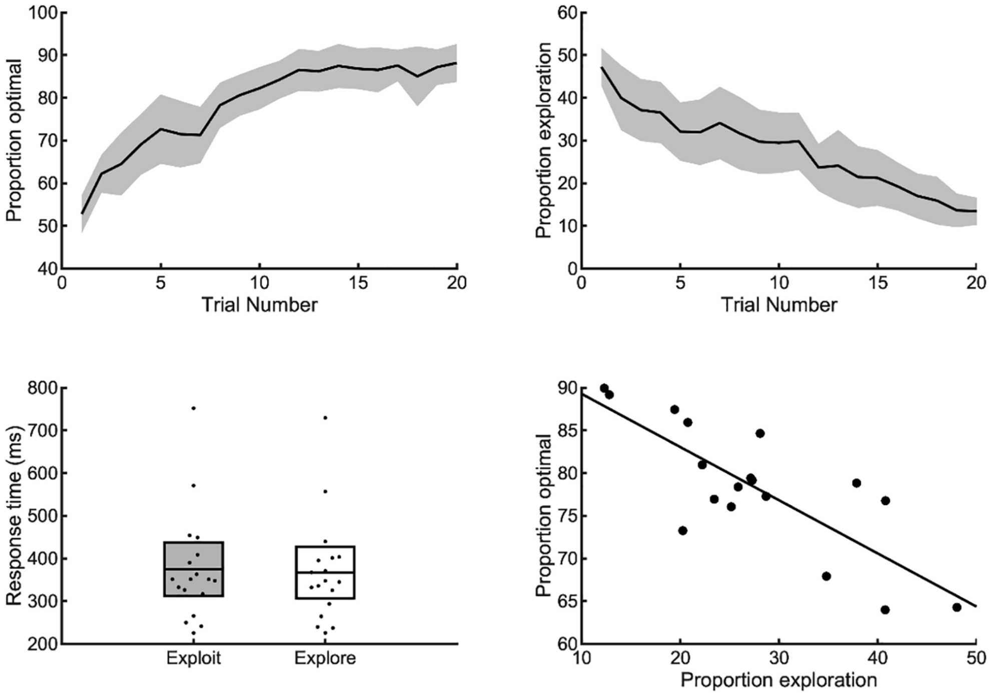

Fig. 2. Behavioral results. Accuracy (top left) was defined as the proportion of trials that participants selected the higher-valued option. A computational model classified each trial as an exploration or exploitation (mean exploration proportion: top left). Mean response time did not differ by decision type (bottom left, individual means also shown). In this task, participants who explored more tended to perform worse (bottom right). The shaded regions and error bars show 95% confidence intervals.

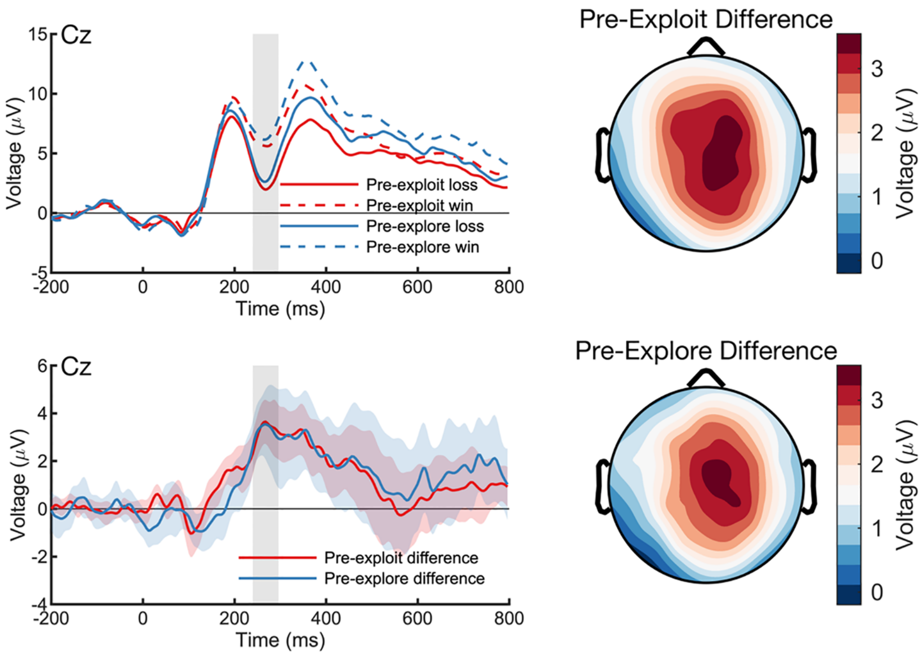

Fig. 3. Reward positivity preceding decisions to exploit and explore. Conditional waveforms are show in the top left panel, and difference waves (win minus loss) are shown in the bottom left panel. The vertical shaded rectangle indicates the analysis window. The shaded regions around each difference wave reflect 95% confidence intervals.

2.6. *Choice-Locked N200*

The choice-locked N200 magnitude was 4.36 μV, 95% CI [2.92, 5.79] on exploitation trials and 3.53 μV, 96% CI [2.02, 5.04] on exploration trials. A paired-sampled *t*-test indicated a statistically-significant difference, $t(17) = -3.24$, $p = .005$, Cohen's $d = -0.76$ (Figs. 5 and 6).

2.7. *Choice-Locked P300*

The bandit-locked P300 magnitude was 5.06 μV, 95% CI [3.76, 6.36] on exploitation trial and 5.26 μV, 96% CI [3.84, 6.68]. A paired-sampled *t*-test indicated no statistically-significant difference, $t(17) = 0.73$, $p = .47$, Cohen's $d = 0.17$ (Fig. 5).

3. Discussion

Our results suggest the involvement of two neural systems when transitioning from an exploitative to an exploratory mode of decision-making. First, feedback-locked phasic activity of the LC-NE system is associated with decisions to explore. Second, exploratory decisions may elicit enhanced response conflict, processed within ACC. These two neural systems – phasic LC-NE activity and conflict-related ACC activity – are indexed by enhancements to the P300 and N200 ERP components, respectively.

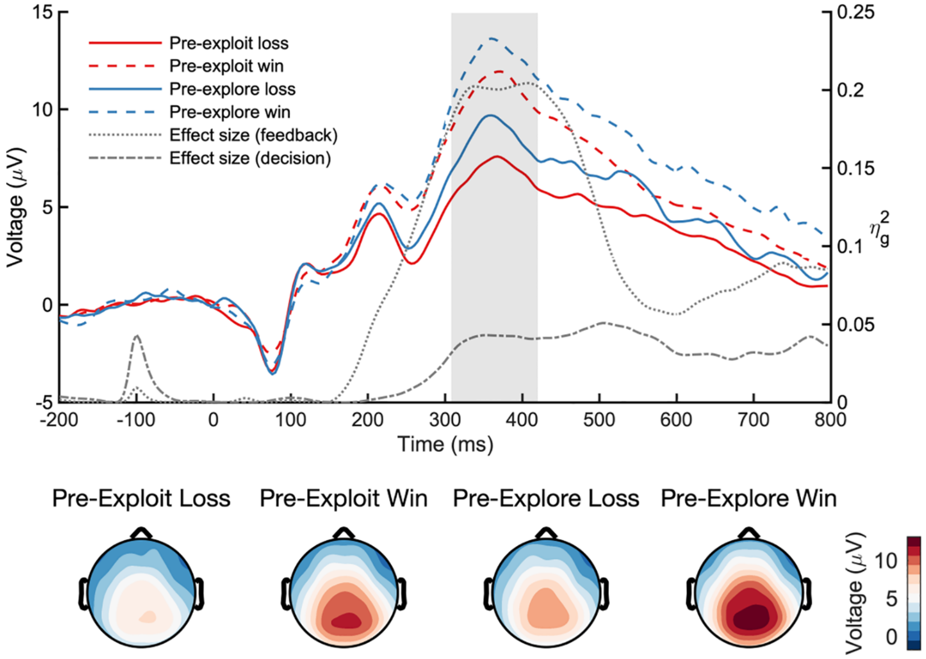

Fig. 4. Feedback-locked P300 waveforms and scalp distributions preceding decisions to exploit and explore. The shaded rectangle indicates the analysis window. The grey lines show effect size ($\eta_g{}^2$) for each main effect (feedback: loss/win, decision: exploit/explore) computed on a moving mean (window length: 100 ms).

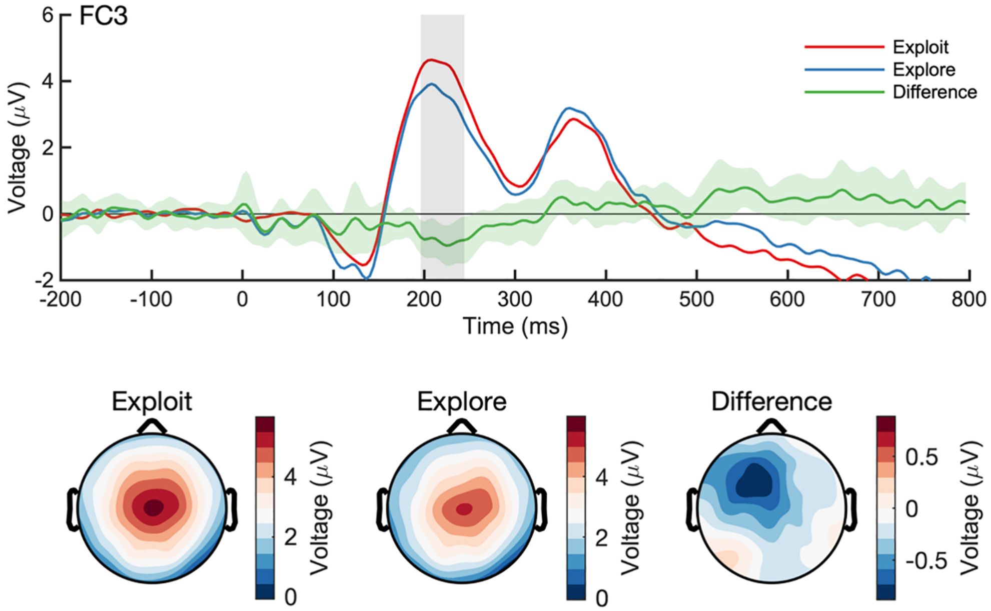

Fig. 5. Choice-locked N200 waveform and scalp distributions. The vertical shaded rectangle indicates the analysis window. The shaded region around the difference wave reflects a 95% confidence interval.

Behaviorally, our participants learned to pick the optimal bandit in a two-armed bandit task. A model, fit to each individual's behavior, determined which responses were exploitations and which were explorations – that is, on which trials the higher-valued bandit was chosen, and on which trials the alternative was chosen. Our task was stationary – outcome contingencies never changed within a block – and more exploration was associated with poorer performance (see correlation in Fig. 2). In general, exploration rate is driven by a combination of individual differences (e.g. Frank et al., 2009) and context. Sequential decision problems, such as bandit tasks, may be stationary or non-stationary, and may have any number of choices/actions. A full review of the decision-making literature is beyond the scope of this study, but it is worth mentioning a couple of relevant examples. Jepma and Nieuwenhuis (2011) used a four-armed bandit with continuously drifting average rewards, similar to Daw et al. (2006). They reported a mean exploration rate of 31%. Blanchard and Gershman (2018) used a two-armed bandit with only occasional reversals (5% of trials). Their participants' average exploration rate dropped from 85% to 12% over a 50-trial block. The point of these examples is to illustrate that exploration rate can vary greatly across experiments, and that our mean exploration rate – which dropped from 47% to 13% over a 20-trial block – is in line with previous bandit studies (Fig. 2).

3.1. Neural response to feedback

By categorizing choices as either explorations or exploitations, we were able to examine the neural response to feedback preceding each decision type. We observed an ERP difference at a scalp location and time range consistent with the P300, an ERP component that, like the N200, is thought to relate to a phasic release of norepinephrine (Nieuwenhuis et al., 2005). The neurotransmitter norepinephrine has featured heavily in several theories of decision making. Relevant here is the view that norepinephrine indexes a neural interrupt – a signal that one's current model of the world might be erroneous, potentially requiring a strategy switch (Bouret and Sara, 2005; Dayan and Yu, 2006; Yu and Dayan, 2005). For example, imagine a participant in the current study trying to win as many points as possible by selecting which of two slot machines to play (the two-armed bandit task). The participant is told that one of the bandits yields a greater average reward than the other. Initially, the participant continues to select one of the bandits because the payouts seem high. At some point, however, the participant decides explore the other option. We argue that this switch – from deciding to exploit one option, to deciding to explore the other – is one example of the neural interrupt discussed by (Bouret and Sara, 2005; Dayan and Yu, 2006; Yu and Dayan, 2005). Supporting this assertion is our observation that feedback preceding decisions to explore (and less-

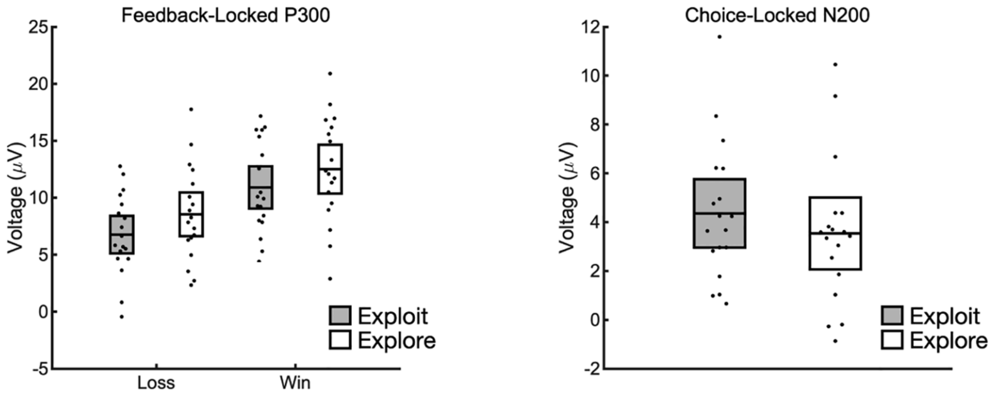

Fig. 6. Summary of results. There was a main effect of upcoming decision type on the feedback-locked P300 (left) and choice-locked N200 (right). Error bars show the 95% confidence intervals.

likely decisions, in general) elicited an enhanced P300 compared to feedback preceding decisions to exploit.

Our other feedback-related hypothesis involved the reward positivity, an ERP component thought to index the phasic release of dopamine (Holroyd and Coles, 2002). Based on previous research linking tonic dopamine and exploration (and other work suggesting that tonic dopamine effects the reward positivity) we hypothesized an effect of decision type (exploit/explore) on the reward positivity. However, although we observed a robust reward positivity for both exploitations and explorations, we found no statistically-significant effect of decision type (Fig. 3). One possible confound here is component overlap. Because the P300 and reward positivity components overlap in time, examining the reward positivity is problematic when P300 effects are also present (such as in the present study – see Figs. 3 and 4). P300 contamination is usually due to frequency effects (e.g., when losses are less frequent than wins) but it's possible that other events affecting the P300 – such as the neural processes associated with exploration – may have hindered our reward positivity investigation. We suggest that to properly examine the role of the reward positivity in the explore/exploit trade-off would require a task for which the P300 is unaffected by feedback valence (as it is here). See Krigolson (2017) for a discussion of component overlap and other methodological considerations.

3.2. *Neural response to bandits*

Following feedback, participants were presented with the choice stimuli again, i.e. the bandits. We observed that the choice-locked N200 ERP component was greater for explorations compared to exploitations. Note, however, that this analysis was exploratory – the observed N200 effect was only apparent after an examination of the difference waveform, and appeared to overlap with a P200 component. Furthermore, we observed the predicted enhancement of the bandit-locked N200 prior to explorations, but no enhancement of the bandit-locked P300. These components often co-occur and are referred to as the N200/P300 (or N2/P3) complex (Duncan-Johnson and Donchin, 1977). According the modified LC-P3 theory, both the N200 and the P300 depend on phasic norepinephrine (Hong et al., 2014; Mückschel et al., 2017; Warren and Holroyd, 2012; Warren et al., 2011). In particular, the modified LC-P3 theory suggests that phasic bursts of norepinephrine have two effects: an initial cortical enhancement between 200 and 300 ms, and a later cortical impairment between 300 and 600 ms. In other words, phasic norepinephrine enhances both the N200 and the P300, but through different mechanisms (N200: abundance, P300: depletion). To be consistent with the modified LC-P3 theory, we must conclude that our bandit stimuli did not elicit a greater phasic release of norepinephrine prior to decisions to explore compared to decisions to exploit. If they had, we would have observed an exploration-dependent enhancement of both the N200 and the P300. We are thus left with the following question: what could elicit an enhancement of the N200 but not the P300?

To answer this question, we turn to the cognitive control and conflict monitoring literature. Cognitive control is a set of processes that enable humans to flexibly adapt to new situations and goals. According to the conflict-monitoring hypothesis, the need for cognitive control is triggered via the detection of information processing conflict (Botvinick et al., 2001). For example, incongruent stimuli in the Stroop task activate two competing responses – reading the word and naming the color – thus eliciting response conflict and a need for control (Botvinick et al., 2001; Stroop, 1935). In the brain, conflict is processed within the ACC, which generates a conflict-dependent N200; incongruent stimuli in a flanker task elicit an enhanced N200 relative to congruent stimuli (Yeung et al., 2004). Tasks that elicit a conflict-dependent N200 tend to also elicit a P300, but Enriquez-Geppert et al. (2010) showed that the N200 mostly indexes conflict, while the P300 mostly indexes motor inhibition. Thus, an N200 effect in the absence of a P300 effect is possible provided that there is response conflict but not motor inhibition.

We speculate that our exploration trials prompted response conflict because of the simultaneous activation of two responses: the computationally valuable exploitative option, and the computationally less valuable exploratory option. Here, exploitations represented the prepotent response and, like go trials in a go/no-go task, elicited low response conflict. Thus, the bandit-locked N200 was enhanced for explorations (high conflict) relative to exploitations (low conflict). We observed no such enhancement of the bandit-locked P300, however. As Enriquez-Geppert et al. (2010) showed, this pattern of results is possible for tasks that elicit response conflict but not motor inhibition. Since motor inhibition is presumably most relevant around the time of the response, this seems a reasonable characterization of our bandit-locked results; our participants were not cued to respond until around one second after the appearance of the bandits. Thus, the appearance of our bandits impacted the N200 but not the P300.

A response-conflict interpretation of our bandit-locked N200 result aligns with work suggesting that the ACC (the neural generator of the conflict-dependent N200) is involved with decisions to explore or exploit only insofar as it is more active during difficult choices. Shenhav et al. (2014) pointed out that foraging experiments tend to confound foraging value – the value associated with exploration – with choice difficulty (i.e., conflict). As the value of switching approaches the value of staying, and exploration becomes more likely, choice difficulty increases. When foraging value and choice difficulty are dissociated, ACC activity tends to track the latter (Shenhav et al., 2014). It is therefore problematic to conclude that the ACC has a special role in foraging beyond the processing of choice difficulty (e.g., in tracking foraging value: Kolling et al., 2012). The enhanced N200 we observed just prior to decisions to explore is consistent with the view that the ACC processes choice difficulty during explore/exploit decisions. It may also be consistent with the view that the ACC processes foraging value (Kolling et al., 2012), as we did not dissociate foraging value and choice difficulty. However, a foraging-value account of our N200 data does not seem as promising as a conflict-monitoring account given the amount of literature linking the ACC-generated N200 to response conflict (Baker and Holroyd, 2011; Enriquez-Geppert et al., 2010; Nieuwenhuis et al., 2003; Yeung et al., 2004).

4. Conclusions

By examining ERPs to two events – feedback and choice stimuli - we demonstrate the contribution of three neural systems to the explore-exploit dilemma. First, phasic activity of the LC-NE system, as indexed by a feedback-locked P300, plays a critical role in triggering a switch from exploitative to explorative decision making. Conversely, phasic midbrain dopamine does not appear to play this same role; the reward positivity, a dopamine-driven RL signal, did not predict decision type. Finally, the period just prior to a decision to explore appears to involve response conflict; the bandit-locked N200, a neural conflict signal originating in ACC, was enhanced prior to exploratory decisions.

5. Experimental procedure

5.1. *Participants*

Twenty-three university-aged participants (9 male, 1 left-handed, M_{age} = 22, 95% CI [21, 23] with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. Participants who did not meet a pre-set accuracy threshold of 60% (as defined below in the Data Analysis subsection) were excluded from the analysis. In total, five participants were excluded from the analysis due to poor performance (mean accuracies of 56%, 51%, 50%, 48%, and 50%). All of the participants were volunteers who received credit in an undergraduate course for their participation. The participants provided informed consent approved by the Health Sciences Research Ethics Board at Dalhousie University.

5.2. Apparatus and procedure

Participants were seated comfortably 75 cm in front of a computer display and used a standard USB keyboard to perform a computerized gambling task, written in MATLAB (Version 7.14, Mathworks, Natick, USA) using the Psychophysics Toolbox Extension (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Participants received both verbal and written instructions and were encouraged to maintain a central fixation and to minimize head movements and eye blinks. Participants were told that the goal of the task was to win as many points as possible.

The experimental task was a two-choice gambling game (i.e., a two-armed bandit: Sutton and Barto, 1998). Comprised of two one-armed bandits, or slot machines, our two-armed bandit required that participants choose between one of two possible gambles, represented by colored squares presented to the left and right of a central fixation point. Participants were told beforehand that one of the choices had a higher average win payout than the other. The loss amounts associated with each choice were equivalent. Thus, participants were gambling based on the win payouts; the overall proportion of wins to losses was not task relevant. At the beginning of a trial, a 1.1 cm fixation cross subtending 0.84 degrees of visual angle was presented for 900–1100 ms. Subsequent to this, two colored squares appeared, each 2.8 cm across and subtending 2.14 degrees of visual angle, equidistant on either side of the fixation cross. The squares were 11.3 cm apart, center-to-center, or 8.62 degrees of visual angle. After the squares were presented for 900–1100 ms, the fixation cross changed color to cue participants to respond by selecting either the left square ('a' key) or right square ('l' key). If participants responded too early, points were deducted from the total won. Similarly, participants were told that points would be deducted if they responded too late (after two seconds). This ensured that all valid responses occurred within a 0–2000 ms window following the go cue.

After a valid response, the squares were removed, leaving only a fixation cross on the display for 400–600 ms. Participants then viewed a feedback stimulus indicating the amount of points won or lost on that trial for 1000 ms. As explained to participants in the instructions, half of the trials resulted in a win, and half of the trials in a loss; specific outcomes were determined by a pseudorandom number generator. Thus, the chance of winning any given trial was 50%. This ensured that a similar number of win and loss trials would be available for later analysis (Holroyd and Krigolson, 2007; Krigolson, 2017). Loss trials resulted in a 5-point deduction, regardless of which square was selected. Wins, on the other hand, always paid a positive amount that was dependent on which square was selected. Each square, or bandit, paid amounts selected from Gaussian distributions with identical variances ($\sigma^2 = 1$), but with different means. A block consisted of 20 gambles, or trials, after which a new block began with two new random colors and two new reward distributions. Participants were told that after each block the squares reset (new colors and payouts), and that they then had to relearn which square was the higher-paying choice in order to win as much as possible. Participants completed 50 blocks in total and were given a self-paced rest break every 10 blocks.

To ensure that the task presented a similar level of difficulty for all participants, payout distributions were initially quite different (means of 6 and 12 points). The mean of the lower valued square was increased by one after every block as long as participants were able to achieve an accuracy of 80% (defined as selecting the higher valued option in the second half of a block at least 8 times out of 10). The payout distributions were fixed once participant accuracy dropped below 80%, i.e. once an appropriate level of difficulty for that participant was achieved. See Fig. 1 for timing details.

5.3. Data collection

The experimental program recorded participant choice (higher or lower valued square) and response time. The EEG was recorded from 64 electrode locations using Brain Vision Recorder software (Version 1.20, Brain Products, GmbH, Munich, Germany). The electrodes were mounted in a fitted cap with a standard 10–20 layout and were recorded with an average reference built into the amplifier. The vertical and horizontal electrooculograms were recorded from electrodes placed above and below the right eye and on the outer canthi of the left and right eyes. Electrode impedances were kept below 20 kΩ. The EEG data were sampled at 1000 Hz and amplified (Quick Amp, Brainproducts, GmbH, Munich, Germany).

5.4. Computational model

Our analysis depended on classifying participant decisions as either exploitations or explorations. To achieve this, we modeled each participant's responses, trial by trial. Our model, used previously in Krigolson et al. (2013), maintained a value for each bandit stimulus on each trial: $v_t(1)$ and $v_t(2)$. The probability on trial t of selecting stimulus i (that is, the likelihood of making an action a_i) was computed as per the *softmax* equation:

$$P_t(a_i) = \frac{e^{v_t(i)/\tau}}{e^{v_t(1)/\tau} + e^{v_t(2)/\tau}}$$

where τ (temperature) determined the degree of bias towards choosing high-valued stimuli (greater bias for lower τ). On each "win" trial, following feedback R_t, a prediction error δ_t was generated for the selected stimulus s according to:

$$\delta_t = R_t - v_t(s)$$

The value of the chosen bandit s was then updated using the following learning rule:

$$v_{t+1}(s) = v_t(s) + \alpha\delta_t$$

in which prediction errors were scaled by α. The value of the unselected stimulus was unchanged. Losses, designed to be uninformative in this task, did not result in any prediction error computation or model update. (Recall that losses occurred with 50% probability, regardless of bandit choice, and only ever resulted in a loss of 5 points). To support this design choice, we compared our model to one in which both wins and losses prompted model updates.

The temperature and learning rate were tuned for each participant. These parameters (τ, α) were tuned using the MATLAB function fmincon (Optimization Toolbox, Release 2018a, Mathworks, Natick). Specifically, we constructed an objective function (the function to be minimized) as the negative log-likelihood of a participant's set of responses. Log-likelihood was computated as:

$$\sum_t \log(P_t(a_s))$$

where $P_t(a_s)$ was the softmax probability associated with the selected bandit s on trial t.

To reiterate: model tuning was done for each participant. Thus, the model-tuning procedure generated learning parameters (τ, α) for each participant. Additionally, we classified trials as exploitations or explorations using the softmax result on each trial. Trials on which the participant made the less likely response, according to the softmax equation, were classified as explorations. All other trials – trials in which the higher-probability response was made – were classified as exploitations. As expected, there were more explorations early in learning (Fig. 2). These trial classifications (exploit/explore) were used to drive the ERP analysis – see below.

5.5. Data analysis

5.5.1. Behavioral

For each participant and trial (1–20) we computed the mean proportion of times, across all blocks, that the optimal choice was made (i.e., the higher-valued bandit was chosen). We also computed the mean of this proportion across all trials and participants. Similarly, we computed the mean proportion of explorations for each trial (1–20) from each participant's trained model. We then computed the average and standard deviation of this exploration proportion across all participants. Finally, we computed the mean response times for each decision type (exploit/explore).

5.5.2. EEG

EEG data were downsampled to 250 Hz, filtered through a (0.1 Hz–30 Hz pass band) phase shift-free Butterworth filter (60 Hz notch), and re-referenced to the average of the two mastoid channels. Next, ocular artifacts were removed using independent component analysis. Subsequent to this, and for each event of interest (stimulus and feedback presentation), 800 ms epochs of EEG data were constructed from 200 ms prior to 800 ms following event onset. All trials were then baseline corrected using a 200 ms pre-event window. Finally, trials in which the change in voltage in any channel exceeded 10 μV per sampling point or the change in voltage across the epoch was greater than 100 μV were discarded. On average, we removed 6.3% of the stimulus-locked epochs (95% CI [3.8, 8.9]) and 6.6% of the feedback-locked trials (95% CI [3.7, 9.5]). Our hypothesis concerned two events: feedback given just prior to a decision to exploit/explore (trial N-1), and the choice stimuli that are exploited/explored (trial N). Below we describe how we quantified ERPs for these two events.

5.5.3. Feedback-Locked ERPs

To quantify the reward positivity, we averaged the feedback-locked EEG for each participant, channel, feedback condition (win/loss), and decision type (exploit/explore). We then constructed difference waveforms by subtracting the average loss waveforms from the average win waveforms (Krigolson, 2017). To identify a window of analysis, we constructed a "grand-grand" average difference waveform (Kappenman and Luck, 2016) by collapsing across both participant and decision type (exploit/explore). We then identified a window of interest by locating the peak of this difference waveform (maximum voltage, across all timepoints and scalp locations), and chose as a half-interval the time on the leading edge of the peak at which 75% of the maximum voltage was reached. Thus, the reward positivity was defined as the mean voltage from 240 to 296 ms post feedback at electrode Cz (See Fig. 3). A reward positivity score was computed for each participant and decision type (pre-exploit/pre-explore). A similar procedure was followed for the P300, except that the grand-grand average also collapsed across feedback type (i.e., it was the average response to all feedback). The peak of the P300 was defined as the maximum positive deflection, across all timepoints and scalp locations, and the half-interval was defined as the point on the leading edge of the waveform at which 75% of the maximum voltage was reached. This resulted in a P300 defined as the mean voltage from 308 to 420 ms post feedback at electrode POz (see Fig. 4). Thus, a P300 score was computed for each participant, feedback type (win/loss), and decision type (pre-exploit/pre-explore).

5.5.4. Choice-Locked ERPs

Preceding a decision to exploit or explore, participants were shown the choice stimuli, or bandits. To analyze the ERPs locked to the bandits, we averaged the choice-locked EEG for each channel and decision type (exploit/explore), for each participant. Only trials with valid behavioral responses were included. To identify the P300 time range, we followed a similar procedure as for the feedback-locked analysis; we collapsed across all participants and conditions (exploit/explore) and found the time/location of greatest voltage. We then took 75% of the leading edge as the half interval. This resulted in a P300 defined as the mean voltage from 312 to 392 ms post bandits at electrode POz (i.e., a P300 score for each participant and decision type). Our N200 analysis was exploratory, as there was no obvious N200 peak at any anterior electrode site that could be identified when we collapsed across decision type. Instead, we identified the time/location of the greatest difference between our average explore waveform and our average exploit waveform. An interval from 196 to 244 ms post bandits at electrode FC3 was identified as the time/location of greatest difference (i.e., where the 95% confidence intervals of the difference wave did not overlap with zero). There are two caveats to this exploratory analysis. First, our N200 definition was biased because it was defined using our conditions of interest (exploit/explore). Second, this time range overlapped with a centrally-located P200 ERP (although the difference was maximal at a frontal site – see Fig. 5).

5.5.5. Single-Trial analysis

To further investigate the relationship between the feedback-locked P300 and an upcoming decision to exploit or explore, we computed a single-trial EEG analysis. A P300 score was generated for each participant and trial using the same procedure as in our feedback-locked ERP analysis (i.e., we averaged the post-feedback voltage from 308 to 420 ms at electrode POz). We then calculated, for each participant, a regression line relating the trial-by-trial P300 score to the softmax probability of the upcoming trial decision. If exploration is associated with greater P300 scores, then we ought to see a negative relationship between P300 magnitude and softmax probability. (Recall that we defined exploration as a decision with a less-than-maximal softmax probability; thus, less-likely decisions ought to be preceded by larger P300s).

5.5.6. Inferential statistics

The existence of a reward positivity, defined as a difference score, was tested using a single-sample *t*-test (Holroyd and Krigolson, 2007; Rodríguez-Fornells, et al., 2002). Between decision conditions (pre-exploitation, pre-exploration) the reward positivity's were compared using a paired samples *t*-test. The feedback-locked P300 was analyzed using a 2 (feedback: win, loss) by 2 (decision type: pre-exploit, pre-explore) repeated-measures ANOVA. The choice-locked N200 scores were compared using a paired-samples *t*-test, as were the choice-locked P300 scores. Finally, participant slopes in our single-trial analysis were compared against zero with a single-sample *t*-test. For all t-tests, we computed Cohen's *d* according to:

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

where M_{diff} was the difference score mean and s_{diff} was the difference score standard deviation (or in the case of the reward positivity, the mean and standard deviation of the ERP score itself; see Cumming, 2014). For the ANOVA, we computed two different effect-size measures: η_p^2 and η_g^2 (Lakens, 2013; Olejnik and Algina, 2003).

Acknowledgements

This research was supported by Natural Sciences and Engineering Research Council of Canada graduate funding (first author) and Discovery and Infrastructure grants (third author).

Appendix

See Fig. A1.

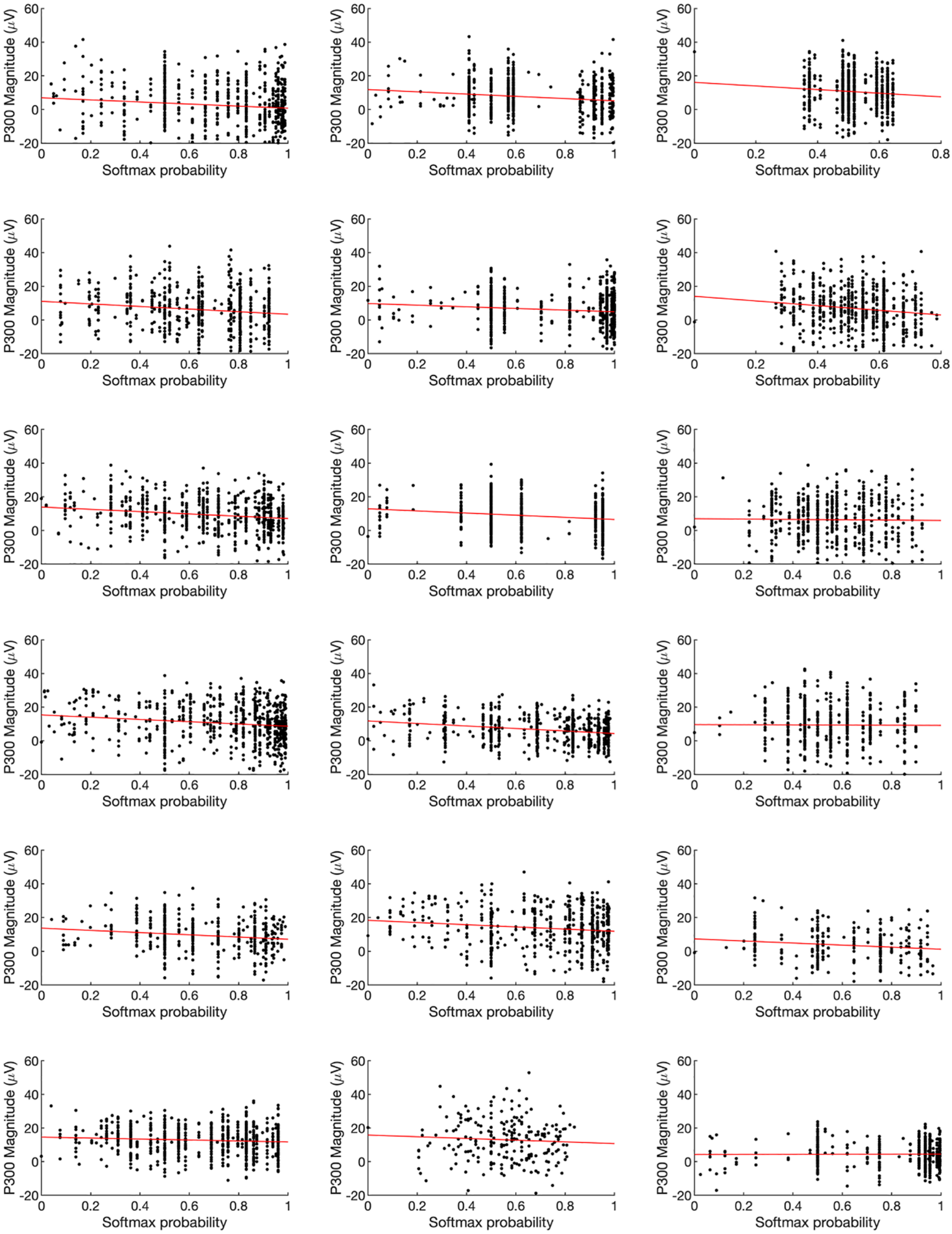

Fig. A1. Relationship between each participant's trial-to-trial P300 and the model-generated likelihood (softmax) of the upcoming decision.

References

Aston-Jones, G., Cohen, J.D., 2005. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu. Rev. Neurosci. 28, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709.

Aston-Jones, G., Rajkowski, J., Kubiak, P., 1997. Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. Neuroscience 80 (3), 697–715.

Baker, T.E., Holroyd, C.B., 2011. Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. Biol. Psychol. 87 (1), 25–34. https://doi.org/10.1016/j.biopsycho.2011.01.010.

Beeler, J.A., 2012. Thorndike's Law 2.0: dopamine and the regulation of thrift. Front. Neurosci. 6. https://doi.org/10.3389/fnins.2012.00116.

Beeler, J.A., Daw, N.D., Frazier, C.R.M., Zhuang, X., 2010. Tonic dopamine modulates exploitation of reward learning. Front. Behav. Neurosci. 4.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10 (9), 1214–1221. https://doi.org/10.1038/nn1954.

Blanchard, T.C., Gershman, S.J., 2018. Pure correlates of exploration and exploitation in the human brain. Cognitive, Affective, Behav. Neurosci. 18 (1), 117–126. https://doi.org/10.3758/s13415-017-0556-2.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D., 2001. Conflict monitoring and cognitive control. Psychol. Rev. 108 (3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624.

Bourdaud, N., Chavarriaga, R., Galan, F., Millan, J.d.R., 2008. Characterizing the EEG correlates of exploratory behavior. IEEE Trans. Neural Syst. Rehabil. Eng. 16 (6), 549–556. https://doi.org/10.1109/TNSRE.2008.926712.

Bouret, S., Sara, S.J., 2005. Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. Trends Neurosci. 28 (11), 574–582. https://doi.org/10.1016/j.tins.2005.09.002.

Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.

Cavanagh, J.F., Figueroa, C.M., Cohen, M.X., Frank, M.J., 2011. Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. Cereb. Cortex. https://doi.org/10.1093/cercor/bhr332. bhr332.

Clayton, E.C., Rajkowski, J., Cohen, J.D., Aston-Jones, G., 2004. Phasic activation of monkey locus ceruleus neurons by simple decisions in a forced-choice task. J. Neurosci. 24 (44), 9914–9920. https://doi.org/10.1523/JNEUROSCI.2446-04.2004.

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

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

Dayan, P., Yu, A.J., 2006. Phasic norepinephrine: a neural interrupt signal for unexpected events. Network: Comput. Neural Syst. 17 (4), 335–350. https://doi.org/10.1080/09548980601004024.

Donchin, E., 1981. Surprise!… surprise? Psychophysiology 18 (5), 493–513. https://doi.org/10.1111/j.1469-8986.1981.tb01815.x.

Duncan-Johnson, C.C., Donchin, E., 1977. On quantifying surprise: the variation of event-related potentials with subjective probability. Psychophysiology 14 (5), 456–467. https://doi.org/10.1111/j.1469-8986.1977.tb01312.x.

Enriquez-Geppert, S., Konrad, C., Pantev, C., Huster, R.J., 2010. Conflict and inhibition differentially affect the N200/P300 complex in a combined go/nogo and stop-signal task. NeuroImage 51 (2), 877–887. https://doi.org/10.1016/j.neuroimage.2010.02.043.

Foti, D., Hajcak, G., 2012. Genetic variation in dopamine moderates neural response during reward anticipation and delivery: evidence from event-related potentials. Psychophysiology 49 (5), 617–626. https://doi.org/10.1111/j.1469-8986.2011.01343.x.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., Moreno, F., 2009. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat. Neurosci. 12 (8), 1062–1068. https://doi.org/10.1038/nn.2342.

Gilzenrat, M.S., Nieuwenhuis, S., Jepma, M., Cohen, J.D., 2010. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. Cognitive, Affective Behav. Neurosci. 10 (2), 252–269. https://doi.org/10.3758/CABN.10.2.252.

Gittins, J., Jones, D., 1974. A dynamic allocation index for the sequential design of experiments. In: Gani, J. (Ed.), Progress in Statistics, pp. 241–266 North-Holland.

Grace, A.A., Floresco, S.B., Goto, Y., Lodge, D.J., 2007. Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. Trends Neurosci. 30 (5), 220–227. https://doi.org/10.1016/j.tins.2007.03.003.

Hassall, C.D., Holland, K., Krigolson, O.E., 2013. What do I do now? An electroencephalographic investigation of the explore/exploit dilemma. Neuroscience 228, 361–370. https://doi.org/10.1016/j.neuroscience.2012.10.040.

Heekeren, H.R., Marrett, S., Ungerleider, L.G., 2008. The neural systems that mediate human perceptual decision making. Nat. Rev. Neurosci. 9 (6), 467–479. https://doi.org/10.1038/nrn2374.

Holroyd, C.B., Coles, M.G.H., 2002. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109 (4), 679–709. https://doi.org/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 (6), 913–917. https://doi.org/10.1111/j.1469-8986.2007.00561.x.

Holroyd, C.B., McClure, S.M., 2015. Hierarchical control over effortful behavior by rodent medial frontal cortex: a computational model. Psychol. Rev. 122 (1), 54–83. https://doi.org/10.1037/a0038339.

Holroyd, C.B., Yeung, N., 2012. Motivation of extended behaviors by anterior cingulate cortex. Trends Cognitive Sci. 16 (2), 122–128. https://doi.org/10.1016/j.tics.2011.12.008.

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

Hong, L., Walz, J.M., Sajda, P., 2014. Your eyes give you away: prestimulus changes in pupil diameter correlate with poststimulus task-related EEG dynamics. PLoS One 9 (3), e91321. https://doi.org/10.1371/journal.pone.0091321.

Jepma, M., Nieuwenhuis, S., 2011. Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. J. Cognit. Neurosci. 23 (7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548.

Kane, G.A., Vazey, E.M., Wilson, R.C., Shenhav, A., Daw, N.D., Aston-Jones, G., Cohen, J.D., 2017. Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. Cognitive, Affective, Behav. Neurosci. 17 (6), 1073–1083. https://doi.org/10.3758/s13415-017-0531-y.

Kappenman, E.S., Luck, S.J., 2016. Best practices for event-related potential research in clinical populations. Biol. Psychiatry: Cognitive Neurosci. Neuroimaging 1 (2), 110–115. https://doi.org/10.1016/j.bpsc.2015.11.007.

Kayser, A.S., Mitchell, J.M., Weinstein, D., Frank, M.J., 2015. Dopamine, locus of control, and the exploration-exploitation tradeoff. Neuropsychopharmacology 40 (2), 454–462. https://doi.org/10.1038/npp.2014.193.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., 2007. What's new in psychtoolbox-3. Perception 36 (14), 1.

Kolling, N., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., 2012. Neural mechanisms of foraging. Science 336 (6077), 95–98. https://doi.org/10.1126/SCIENCE.1216930.

Krigolson, O.E., 2017. Event-related brain potentials and the study of reward processing: methodological considerations. Int. J. Psychophysiol. https://doi.org/10.1016/j.ijpsycho.2017.11.007.

Krigolson, O.E., Hassall, C.D., Handy, T.C., 2013. How we learn to make decisions: rapid propagation of reinforcement learning prediction errors in humans. J. Cognit. Neurosci. 26 (3), 635–644. https://doi.org/10.1162/jocn_a_00509.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00863.

Lejuez, C.W., Read, J.P., Kahler, C.W., Richards, J.B., Ramsey, S.E., Stuart, G.L., et al., 2002. Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). J. Exp. Psychol.: Appl. 8 (2), 75–84. https://doi.org/10.1037//1076-898X.8.2.75.

Marco-Pallarés, J., Cucurell, D., Cunillera, T., Krämer, U.M., Càmara, E., Nager, W., et al., 2009. Genetic variability in the dopamine system (Dopamine Receptor D4, Catechol-O-Methyltransferase) modulates neurophysiological responses to gains and losses. Biol. Psychiatry 66 (2), 154–161. https://doi.org/10.1016/j.biopsych.2009.01.006.

Mückschel, M., Chmielewski, W., Ziemssen, T., Beste, C., 2017. The norepinephrine system shows information-content specific properties during cognitive control – evidence from EEG and pupillary responses. NeuroImage 149, 44–52. https://doi.org/10.1016/j.neuroimage.2017.01.036.

Murphy, P.R., Robertson, I.H., Balsters, J.H., O'connell, R.G., 2011. Pupillometry and P3 index the locus coeruleus–noradrenergic arousal function in humans. Psychophysiology 48 (11), 1532–1543. https://doi.org/10.1111/j.1469-8986.

Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., Ridderinkhof, K.R., 2003. Electrophysiological correlates of anterior cingulate function in a go/no-go task: effects of response conflict and trial type frequency. Cognitive, Affective, Behav. Neurosci. 3 (1), 17–26. https://doi.org/10.3758/CABN.3.1.17.

Nieuwenhuis, S., Aston-Jones, G., Cohen, J.D., 2005. Decision making, the P3, and the locus coeruleus–norepinephrine system. Psychol. Bull. 131 (4), 510–532. https://doi.org/10.1037/0033-2909.131.4.510.

Nieuwenhuis, S., De Geus, E.J., Aston-Jones, G., 2011. The anatomical and functional relationship between the P3 and autonomic components of the orienting response: P3 and orienting response. Psychophysiology 48 (2), 162–175. https://doi.org/10.1111/j.1469-8986.2010.01057.x.

Niv, Y., Daw, N.D., Joel, D., Dayan, P., 2007. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology 191 (3), 507–520. https://doi.org/10.1007/s00213-006-0502-4.

Olejnik, S., Algina, J., 2003. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol. Methods 8 (4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434.

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

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

Shenhav, A., Straccia, M.A., Cohen, J.D., Botvinick, M.M., 2014. Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. Nat. Neurosci. 17 (9), 1249. https://doi.org/10.1038/NN.3771.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. J. Exp. Psychol. 18 (6), 643–662. https://doi.org/10.1037/h0054651.

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

Tzovara, A., Murray, M.M., Bourdaud, N., Chavarriaga, R., Millán, J. del R., De Lucia, M., 2012. The timing of exploratory decision-making revealed by single-trial topographic

EEG analyses. NeuroImage 60 (4), 1959–1969. https://doi.org/10.1016/j.neuroimage.2012.01.136.

Warren, C.M., Holroyd, C.B., 2012. The impact of deliberative strategy dissociates ERP components related to conflict processing vs reinforcement learning. Front. Neurosci. 6. https://doi.org/10.3389/fnins.2012.00043.

Warren, C.M., Tanaka, J.W., Holroyd, C.B., 2011. What can topology changes in the oddball N2 reveal about underlying processes? NeuroReport 22 (17), 870. https://doi.org/10.1097/WNR.0b013e32834bbe1f.

Warren, C.M., Wilson, R.C., van der Wee, N.J., Giltay, E.J., van Noorden, M.S., Cohen, J.D., Nieuwenhuis, S., 2017. The effect of atomoxetine on random and directed exploration in humans. PLoS One 12 (4), e0176034. https://doi.org/10.1371/journal.pone.0176034.

Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111 (4), 931–959. https://doi.org/10.1037/0033-295X.111.4.931.

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