Learning Contextual Reward Expectations for Value Adaptation

Francesco Rigoli[1], **Benjamin Chew**[1,2], **Peter Dayan**[3], **and Raymond J. Dolan**[1,2]

Abstract

■ Substantial evidence indicates that subjective value is adapted to the statistics of reward expected within a given temporal context. However, how these contextual expectations are learned is poorly understood. To examine such learning, we exploited a recent observation that participants performing a gambling task adjust their preferences as a function of context. We show that, in the absence of contextual cues providing reward information, an average reward expectation was learned from recent past experience. Learning dependent on contextual cues emerged when two contexts alternated at a fast rate, whereas both cue-independent and cue-dependent forms of learning were apparent when two contexts alternated at a slower rate. Motivated by these behavioral findings, we reanalyzed a previous fMRI data set to probe the neural substrates of learning contextual reward expectations. We observed a form of reward prediction error related to average reward such that, at option presentation, activity in ventral tegmental area/substantia nigra and ventral striatum correlated positively and negatively, respectively, with the actual and predicted value of options. Moreover, an inverse correlation between activity in ventral tegmental area/substantia nigra (but not striatum) and predicted option value was greater in participants showing enhanced choice adaptation to context. The findings help understanding the mechanisms underlying learning of contextual reward expectation. ■

INTRODUCTION

Substantial evidence indicates that subjective values of monetary outcomes are context-dependent. That is, in order for these values to be consistent with the choices participants make between those outcomes, they must be adjusted according to the other rewards available either immediately (Tsetsos et al., 2016; Louie, Glimcher, & Webb, 2015; Louie, LoFaro, Webb, & Glimcher, 2014; Louie, Khaw, & Glimcher, 2013; Soltani, De Martino, & Camerer, 2012; Tsetsos, Chater, & Usher, 2012; Vlaev, Chater, Stewart, & Brown, 2011; Tsetsos, Usher, & Chater, 2010; Stewart, 2009; Johnson & Busemeyer, 2005; Usher & McClelland, 2004; Stewart, Chater, Stott, & Reimers, 2003; Roe, Busemeyer, & Townsend, 2001; Simonson & Tversky, 1992; Huber, Payne, & Puto, 1982; Tversky, 1972) or expected before the options are presented (Rigoli, Friston, & Dolan, 2016; Rigoli, Friston, Martinelli, et al., 2016; Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, & Dolan, 2016; Louie et al., 2014, 2015; Ludvig, Madan, & Spetch, 2014; Kobayashi, de Carvalho, & Schultz, 2010; Rorie, Gao, McClelland, & Newsome, 2010; Padoa-Schioppa, 2009; Stewart, 2009). We recently investigated the latter form of effect (Rigoli, Friston, & Dolan, 2016; Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016) in a decision-making task involving blocks of trials associated with either a low- or high-value context with overlapping distributions. Here, choice behavior was consistent with a hypothesis that the subjective value of identical options was larger in a low-value context compared with a high-value context. This and similar evidence (Louie et al., 2014, 2015; Ludvig et al., 2014; Stewart, 2009) suggests that subjective values are partially rescaled to the reward expected within a given context.

However, in previous studies of temporal adaptation, participants were explicitly informed before the task about the distribution of contextual reward. Such designs enable an analysis of the way that beliefs about contextual reward impact choice but leave open the question of how such beliefs are learned. Here, we investigate this question by analyzing how beliefs about contextual reward are shaped by experience within a context, including learning when there are multiple (and cued) contexts that alternate. One possibility is that participants might ignore contextual cues and only learn a long-run expected rate of reward (Niv, Daw, Joel, & Dayan, 2007). This average reward could then act as a baseline against which the subjective value of an actual reward is adapted. Alternatively, participants might use contextual cues to learn and maintain separate reward expectations for different contexts and rely on these during value adaptation. A final possibility is that reward expectations dependent and independent of cues are both acquired and exert a combined influence on value adaptation.

[1]The Wellcome Trust Centre for Neuroimaging at University College London, [2]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, UK, [3]Gatsby Computational Neuroscience Unit, University College London

© 2017 Massachusetts Institute of Technology

doi:10.1162/jocn_a_01191

Here, we implemented a study design that enabled us to probe learning of contextual reward expectation and its impact on subjective value attribution and choice. First, in a novel behavioral experiment, we analyzed how previous reward experience drives learning of contextual reward. In a second new behavioral experiment, we considered the role of multiple alternating contexts signaled by different cues. We adopted a choice task similar to a previous study (Rigoli, Friston, & Dolan, 2016), but unlike this previous study, in this instance participants were not explicitly instructed about contextual reward distributions and could only learn these distributions observationally by playing the task. The results of these two new experiments provided a motivation for us to examine the neural substrates of learning contextual reward expectations by reanalyzing a previously reported data set (Rigoli, Rutledge, Dayan, et al., 2016) where we used a similar paradigm in conjunction with acquiring fMRI data.

It is well established that, when a reward outcome is presented, neurophysiological and neuroimaging responses in ventral striatum and ventral tegmental area/substantia nigra (VTA/SN) reflect a reward prediction error (RPE) signal (Lak, Stauffer, & Schultz, 2014; Stauffer, Lak, & Schultz, 2014; Niv, Edlund, Dayan, & O'Doherty, 2012; Park et al., 2012; D'Ardenne, McClure, Nystrom, & Cohen, 2008; Tobler, Fiorillo, & Schultz, 2005; O'Doherty et al., 2004; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Schultz, Dayan, & Montague, 1997). This is based on the observation that response in these regions correlates positively and negatively with the actual and expected reward outcome, respectively (Niv et al., 2012; Niv & Schoenbaum, 2008). However, the question remains as to whether these regions also show an RPE signal at the time of presentation of the options rather than the outcomes. In this regard, at option presentation, research has shown a correlation between brain activation and actual option expected value (EV; Lak et al., 2014; Stauffer et al., 2014; Niv et al., 2012; Park et al., 2012; D'Ardenne et al., 2008; Tobler et al., 2005; O'Doherty et al., 2003, 2004; Schultz et al., 1997). However, it remains unknown whether there is also an inverse correlation with the predicted option EV (which is the other component of an RPE signal; Niv et al., 2012; Niv & Schoenbaum, 2008).

These findings motivated a proposal that there might be a distinct effect of outcomes and options on activity in ventral striatum and VTA/SN, corresponding to signaling RPE and EV, respectively (Bartra, McGuire, & Kable, 2013). For example, the possibility that option presentation elicits EV and not RPE signaling is consistent with the idea that expectations about options (which is a key component of the RPE signal) may be fixed or may change over such a long timescale that they would be undetectable within the timescale of an fMRI experiment. However, other theoretical models (Schultz et al., 1997) imply that the VTA/SN (and, by extension, ventral striatum) reflects RPE also at option presentation. This possibility is also consistent with the idea explored here that the value of options is adapted to the context learned from experience, as such context would determine the predicted option EV. Our fMRI analysis aimed to clarify whether presenting options elicits RPE or EV signaling in ventral striatum and VTA/SN and, if RPE is signaled, whether this is related with the effect of context on choice behavior.

METHODS

Participants

Twenty-four healthy, right-handed adults (13 women, aged 20–40 years, mean age = 24 years) participated in the first behavioral experiment. Twenty-eight healthy, right-handed adults participated in the second behavioral experiment. We discarded data from three participants in the second experiment who did not attend properly to the task, as evidenced by having more than 300 (i.e., one half of all) trials with RT shorter than 300 msec (for the other participants, the maximum number of such trials was 37). Therefore, the total sample for the second experiment was 25 participants (15 women, aged 20–40 years, mean age = 25 years). We also reanalyzed data from a previously reported fMRI study where the experimental sample included 21 participants (13 women, aged 20–40 years, mean age = 27 years; for details, see Rigoli, Rutledge, Dayan, et al., 2016). All studies were approved by the University College London research ethics committee.

Experimental Paradigm and Procedure

Participants were tested at the Wellcome Trust Centre for Neuroimaging at University College London. Each experiment involved a computer-based decision-making task lasting approximately 40 min. Before the task, participants were fully instructed about task rules and the basis of payment. Crucially, in Experiments 1 and 2, participants were not informed about the distribution of options that would be encountered during the task (see below). Note that this is a key difference from the tasks adopted in previous studies where participants were instructed about the distributions (Rigoli, Friston, & Dolan, 2016; Rigoli, Rutledge, Dayan, et al., 2016). In the fMRI experiment, before each block, information about the reward distributions was provided (see below).

Experiment 1

On each trial, participants chose between a sure monetary amount, which changed trial by trial (600 trials overall), and a gamble whose prospects were always either zero or double the sure amount, each with equal (50–50) probability (Figure 1A; during instructions, participants were informed about this probability). This ensured that both options always had equal (objective) EV. Trial EV was randomly drawn from a uniform distribution (with

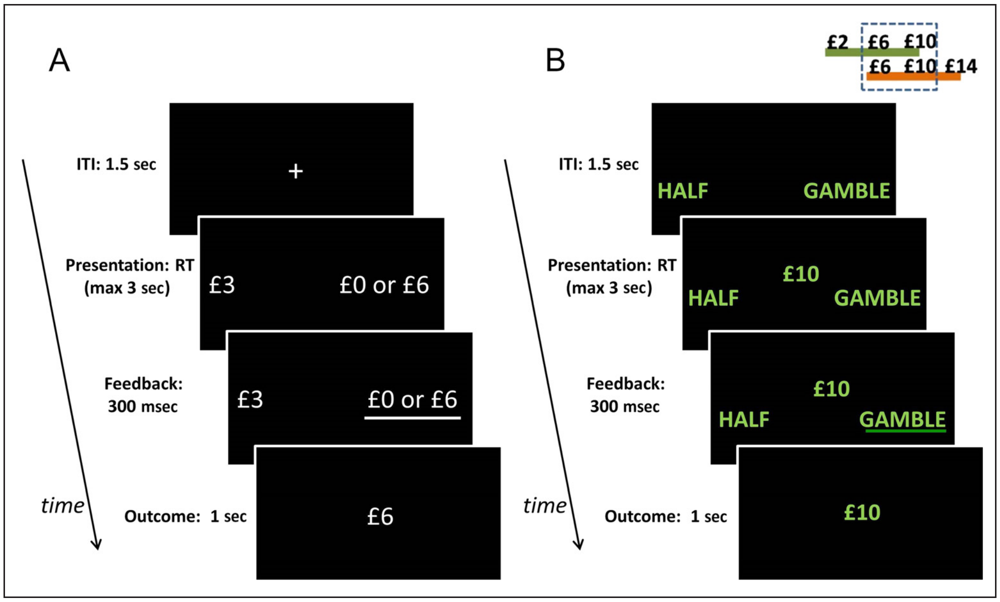


Figure 1. (A) Experimental paradigm for Experiment 1. Participants repeatedly made choices between a sure monetary reward (on the left in the example) and a gamble (on the right in the example) associated with a 50% probability of either double the sure reward or zero. After a decision was performed, the chosen option was underlined, and 300 msec later the trial outcome was shown for 1 sec. The intertrial interval (ITI) was 1.5 sec. At the end of the experiment, a single randomly chosen outcome was paid out to participants. (B) Experimental paradigm for Experiment 2. On each trial, a monetary reward was presented (£10 in the example), and participants had to choose between half of the amount for sure (by pressing the left button) and a 50–50 gamble associated with either the full amount or a zero outcome. A trial started with an ITI lasting 1.5 sec where the two options (i.e., half and gambling) were displayed on the bottom of the screen (on the left and right side, respectively). Next, the trial amount was displayed. Right after a response was performed, the chosen option was underlined for 300 msec, followed by the outcome of the choice, shown for 1 sec. The task was organized in short blocks, each comprising five trials. Each block was associated with one of two contexts that determined the possible EVs within the block. These EVs were £1, £3, and £5 for the low-value context and £3, £5, and £7 for the high-value context. Contexts were signaled by the color of the text on the screen, with low-value context associated with green and high-value context with orange for half of the participants and vice versa for the other half. At the end of the experiment, a single randomly chosen outcome was paid out to participants.

50p steps) in the £1–£6 range. The certain and risky options were presented pseudorandomly on two sides of a screen; participants chose the left or right option by pressing the corresponding button of a keypad. Immediately after a choice was made, the chosen option was underlined for 300 msec and the outcome of the choice was then displayed for one second. Participants had 3 sec to make their choices; otherwise, the statement "too late" appeared, and they received a zero outcome amount. The outcomes of the gamble were pseudorandomized. At the end of the experiment, one trial outcome was randomly selected and added to an initial participation payment of £5. Compared with using the sum of payoffs across all trials, using a single trial for payment minimizes the influence of past outcomes and allowed us to use choices characterized by larger monetary amounts. Because participants do not know ahead of time which trial will be selected, they should work equally hard on each. This is a method of payment routinely used in experimental economics.

This task was used because it has some similarity to the one we used in previous studies (Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016). These studies showed that adaptation to context predisposes participants who prefer to gamble for large EVs to gamble more when EVs are larger relative to contextual expectations and participants who prefer to gamble for small EVs to gamble more when EVs are smaller relative to contextual expectations. Crucially, in our previous studies, contextual expectations were induced descriptively using explicit instructions, whereas here we investigated whether contextual expectations (and the ensuing adaptation effects on choice) arise observationally from option EVs presented on previous trials.

Experiment 2

On each trial, a monetary amount, changing trial by trial (600 trials overall), was presented in the center of the screen, and participants had to choose whether to accept half of it for sure (pressing a left button) or select a gamble whose outcomes were either zero or the amount presented on the screen (i.e., double the sure amount), each with equal (50–50) probability (Figure 1B; during instructions, participants were informed about this probability). As in Experiment 1, this ensured that on every trial the sure option and the gamble always had the same EV. A trial started with an intertrial interval lasting 1.5 sec where the two options (i.e., half and gambling) were displayed on the bottom of the screen (on the left and right side, respectively). Next, the trial amount was displayed. Immediately after a response, the chosen option was underlined for 300 msec, and this was followed by the outcome of the choice, which was shown for 1 sec. Participants had 3 sec to make their choices; otherwise, the statement "too late" appeared, and they received a zero outcome amount.

The task was organized in short blocks, each comprising five trials. Each block was associated with one of two contexts (low value and high value) that determined the possible EVs within the block. These EVs were £1, £3, and £5 for the low-value context and £3, £5, and £7 for the high-value context. Contexts were signaled by the color of the text (green or orange) on the screen, with low-value context associated with green and high-value context with orange for half of the participants and vice versa for the other half. Before a new block started, the statement "New set" appeared for 2 sec. Crucially, during instructions, participants were not told that colors indicated two different reward distributions. The order of blocks, trial amounts, and outcomes were pseudorandomized. At the end of the experiment, one trial was randomly selected among those received, and the outcome that accrued was added to an initial participation payment of £5.

This task was used because it has some similarity to the one we used in a previous study that was successful in eliciting contextual adaptation with contexts that alternated (Rigoli, Friston, & Dolan, 2016). Crucially, in our previous study, contextual expectations were induced descriptively using explicit instructions, whereas here we investigated whether contextual expectations (and the ensuing adaptation effects on choice) arise observationally from experience with cues and/or with option EVs presented on previous trials.

fMRI Experiment

The task was performed inside the scanner (560 trials overall). The design was similar to the task used in Experiment 1, except for two differences (for details, see Rigoli, Rutledge, Dayan, et al., 2016). First, immediately after the choice was made, the chosen option was not underlined, but the unchosen option disappeared for 300 msec. Second, trials were arranged in four blocks (140 trials each). In each block, the sure amount was randomly drawn from a uniform distribution (with 10p steps) within a £1–£5 range (for two blocks: low-value context) or within the £2–£6 range (for the two other blocks: high-value context). Blocks were interleaved with 10-sec breaks. During the interblock interval, a panel showed the reward range associated with the upcoming block. Block order was counterbalanced across participants. At the end of the experiment, one trial was randomly selected among those received, and the outcome that accrued was added to an initial participation payment of £17. Inside the scanner, participants performed the task in two separate sessions, followed by a 12-min structural scan. After scanning, participants were debriefed and informed about their total remuneration.

Please note that the tasks used in Experiments 1 and 2 and the fMRI experiment are different in certain details. For example, the first experiment and the fMRI experiment require choosing a sure monetary amount or a gamble between double the amount or zero, whereas in the second experiment, a monetary amount is presented and participants are asked to choose half of the amount or a gamble between the full amount and zero. However, please note that the differences among experiments do not affect our analyses and results, as our research questions were not based on comparisons between the experiments (see below).

Behavioral Analysis

In all experiments, for analyses we discarded trials where RTs were slower than 3 sec (because it was our time limit followed by the statement "too late") and faster than 300 msec (as this is a standard cutoff for decision tasks; e.g., Ratcliff, Thapar, & McKoon, 2001), resulting in the following average number of trials analyzed per participant: 549 in Experiment 1, 535 in Experiment 2, and 556 in the fMRI experiment. A two-tailed $p < .05$ was employed as significance threshold in all behavioral analyses.

Our main hypothesis was that contextual reward expectations are learned from previous trials and drive choice adaptation. Learning implies that the expected contextual reward at trial t is lower/higher when a low/high EV is presented at trial $t - 1$. Following previous data (Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016), adaptation to context implies that participants who prefer to gamble for large EVs (at trial t) gamble more when EVs are larger relative to contextual expectations, whereas participants who prefer to gamble for small EVs gamble more when EVs are smaller relative to contextual expectations. To assess these predictions, for each participant, we built a logistic regression model of choice (i.e., with dependent measure being choice of the gamble or of the sure option), which included the EV at trial t and the EV at trial $t - 1$ as regressors. Our hypothesis that there would be adaptation to the context (where context is defined simply by the previous trial) predicted an inverse correlation between the effect of EV at trial t and the effect of EV at trial $t - 1$ on gambling percentage. This would indicate that participants who gambled more with larger EVs at trial t would also gamble more with smaller EVs at trial $t - 1$, and participants who gambled more with smaller EVs at trial t would also gamble more with larger EVs at trial $t - 1$.

To probe the computational mechanisms underlying choice behavior, we used computational modeling and performed two distinct analyses. First, we analyzed the influence of the EV at trial t (recall that the sure option and the gamble had equivalent EV) together with the influence of EV at previous trials. To do this, we fitted an exponential decay model to the gambling data, which prescribed that the probability of gambling depends on a sigmoidal function of an intercept parameter β_0 plus a weight parameter β_1 multiplied by the EV at trial t, plus the sum of j weight parameter β_2, each multiplied by the EV at trial $t - j$ and by an exponential decay factor dependent on a parameter λ (which was bounded between 0 and 5 during estimation):

$$P(\text{gambling}) = \sigma\left(\beta_0 + \beta_1 R_t + \beta_2 \sum_{j=1}^{4} e^{-\lambda(j-1)} R_{t-j}\right) \quad (1)$$

Our second modeling analysis consisted in using a computational model that included a learning component, an adaptation component, and a choice component. The learning component establishes that, on every trial, participants update a belief about the expected (i.e., average) EV of options $\bar{r}$ (i.e., contextual reward). A first possibility is that this belief update is based on a delta rule with a learning rate η, which remains constant throughout the whole task (Rescorla & Wagner, 1972). If the EV presented at trial t is R_t (remember that the two options available have equivalent EV), then the EV of options expected at trial $t + 1$ is

$$\bar{r}_{t+1} = \bar{r}_t + \eta(R_t - \bar{r}_t) \quad (2)$$

A second possibility is that the contextual reward expectation is updated following a decreasing learning rate. This can be derived from a Bayesian learning scheme (Bishop, 2006) in which the long-run mean is assumed to be fixed across time (note that a constant learning rate implemented in the model above can be derived from a Bayesian scheme too, in this case assuming a long-run mean, which changes with a constant rate). A decreasing learning rate emerges if we assume, at every trial t, a Gaussian prior distribution with mean $\bar{r}_t$ and precision (i.e., inverse variance) π_t and a new observation (of the EV of options) R_t associated with precision π_R. The posterior (which will correspond to the prior for the next trial) reward expectation corresponds to a prediction error $(R_t - \bar{r}_t)$ multiplied by a learning rate η_t:

$$\bar{r}_{t+1} = \bar{r}_t + \eta_t(R_t - \bar{r}_t) \quad (3)$$

where the learning rate η_t varies on every trial and depends on the two precisions π_R and π_t:

$$\eta_t = \frac{\pi_R}{\pi_t + \pi_R} \quad (4)$$

The posterior precision (which will correspond to the prior precision of the next trial) is equal to:

$$\pi_{t+1} = \pi_t + \pi_R \quad (5)$$

Assuming (as we did in our models with decreasing learning rate) a prior precision at the first trial equal to zero (i.e., $\pi_1 = 0$), Equations 4 and 5 imply that $\eta_t = 1/t$ and hence $\eta_{t+1} < \eta_t$. For instance, the learning rate will be smaller than 0.05 after 20 trials only (formally: $\eta_{t>20} < 0.05$). Note that, in models with decreasing learning rate, the learning rate is not a free parameter. In addition, $\pi_1 = 0$ and Equation 4 imply that the learning rates across trials are independent of the value assigned to π_R (hence, π_R is not a free parameter either, and we set $\pi_R = 1$ in our models).

In Experiment 2 and in the fMRI experiment, where two contexts (signaled by distinct cues) alternate, we analyzed models that considered separate cue-related average reward expectations $\bar{r}_{t,k}$ ($k = 1$ and $k = 2$ for the low- and high-value contexts, respectively; trials for cue k are indexed by t_k). Note that these models assume one separate succession of trials per cue and not that learning is restarted every time a cue appears again. As above, learning could be realized either through a constant or a decreasing learning rate. In addition, for these experiments, we considered models where both a cue-independent $\bar{r}_t$ and a cue-dependent $\bar{r}_{t,k}$ average reward representation were learned in parallel, and both influenced adaptation of incentive value.

The adaptation component of the models is derived from our previous work on value normalization (Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016). Here we extend this by comparing subtractive versus divisive forms of normalization. The adaptation component prescribes that the objective EV of options is transformed into a subjective value by being rescaled to the prevailing average reward. If the objective EV of options at trial t is R_t, then the corresponding subjective value will be

$$V_t(R_t) = R_t - \tau\bar{r}_t \quad (6)$$

In models where $\bar{r}_{t,k}$ (instead of $\bar{r}_t$) is learned, Equation 6 corresponds to $V_t(R_t) = R_t - \tau\bar{r}_{t,k}$. The context parameter τ implements (subtractive) normalization of the objective EV to a degree that is proportional to the average reward $\bar{r}$. We compared this formulation based on subtractive normalization with a model implementing divisive normalization (as suggested, for instance, by some recent neural accounts; Louie et al., 2013, 2014) where R_t is divided by the context parameter τ and the average reward $\bar{r}$:

$$V_t(R_t) = R_t/(1 + \tau\bar{r}_t) \quad (7)$$

In Experiment 2 and in the fMRI experiment, where two contexts alternate, we considered models where adaptation was implemented with respect to both a cue-independent belief about average reward $\bar{r}_t$ and the average reward expected for the current cue $\bar{r}_{t,k}$:

$$V_t(R_t) = R_t - \tau\left(\frac{\bar{r}_t + \bar{r}_{t,k}}{2}\right) \quad (8)$$

The context parameter τ implements (subtractive) normalization associated with both the cue-independent average reward $\bar{r}_t$ and the average reward $\bar{r}_{t,k}$ expected for the current cue k. As above, we also considered a formulation implementing divisive normalization where

$$V_t(R_t) = R_t/\left(1 + \tau\left(\frac{\bar{r}_t + \bar{r}_{t,k}}{2}\right)\right) \quad (9)$$

For some of the models that consider both $\bar{r}_t$ and $\bar{r}_{t,k}$, we implemented separate context parameters, rendering

Equations 8 and 9 $V_t(R_t) = R_t - (\tau_1\bar{r}_t + \tau_2\bar{r}_{t,k})$ and $V_t(R_t) = R_t/(1 + \tau_1\bar{r}_t + \tau_2\bar{r}_{t,k})$, respectively.

Finally, the choice component determines the probability of gambling as determined by a sigmoidal function:

$$P(\text{gambling}) = \sigma(\alpha V_t(R_t) + \mu) \\ = 1/(1 + \exp(-\alpha V_t(R_t) - \mu)) \qquad (10)$$

where α is a value-function parameter, which determines whether gambling is more likely with larger ($\alpha > 0$) or smaller ($\alpha < 0$) subjective value $V(R)$, and μ represents a gambling bias parameter. This implementation of the choice component is motivated by the fact that, in our task and with a linear mapping from objective to subjective values assumed in our models, the sure option and the gamble have equivalent EV, implying that the trial EV is the only variable changing trial-by-trial. This entails that a logistic regression model is sufficient to capture a wide range of mechanistic models of choice (e.g., those based on risk-return accounts; see Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016, for details). The free parameters of the model are the value function parameter α, the gambling bias parameter μ, the context parameter τ, and the learning rate η. The effects postulated by the model (assuming subtractive normalization) in determining gambling probability as a function of different trial EV and different parameter sets are represented in Figure 2A–B. This shows that (i) for positive and negative value function parameter α, the propensity to gamble for larger EVs increases and decreases, respectively; (ii) larger gambling bias parameter μ increases the overall propensity to gamble; (iii) the context parameter τ determines whether, as the estimated average reward $\bar{r}$ increases, the subjective values attributed to EVs increase ($\tau < 0$) or, as predicted by a value normalization hypothesis, decrease ($\tau > 0$) and in so doing exert an impact on gambling propensity; and (iv) the learning rate η determines the extent to which $\bar{r}$ is revised with new experience.

The free parameters were fit to choice data using the *fminsearchbnd* function of the Optimization toolbox in Matlab (see Supplementary Figures S1, S2, and S3 for distributions of parameters). The learning rate η was constrained between 0 and 1, which are the natural boundaries for this parameter. Starting values for parameter estimation was 0 for all parameters. The full models and nested models (where one or more parameters were fixed to 0) were fitted to choice data. For each model, the negative log-likelihood of choice data given the best fitting parameters was computed participant by participant and summed across participants, and the sum of negative log-likelihood was used to compute the Bayesian Information Criterion (BIC) scores (Daw, 2011). These were considered for model comparison, which assigns a higher posterior likelihood to a generative model with smaller BIC.

The value of $\bar{r}_t$ and $\bar{r}_{t,k}$ at the start of the task was set to the true overall average EV across trials (in Experiment 1: £3.5; in Experiment 2: £4; in fMRI experiment: £3.5). To

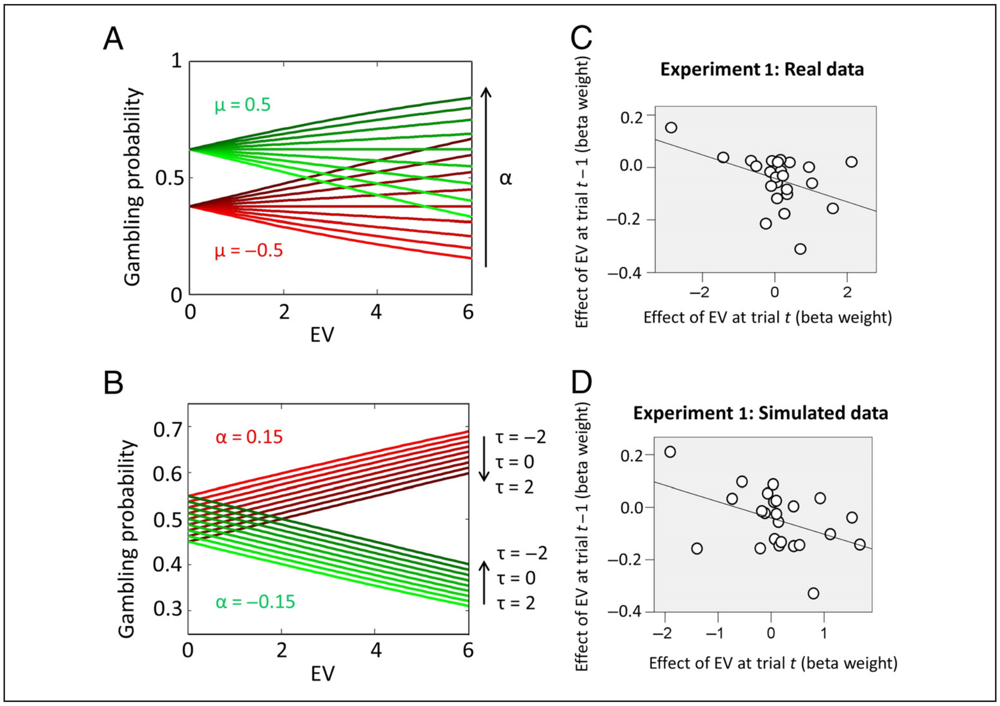


Figure 2. (A) Plots of the gambling probability as a function of trial EV (remember that the two options always had equivalent EV) for agents with specific parameters simulated with the computational model of behavior. Effect of varying the value function parameter α (from −0.2 to 0.2 with increases in 0.05 steps, represented along a bright-to-dark gradient) and the gambling bias parameter μ (green and red lines implement μ = £0.5 and μ = −£0.5, respectively). It is evident that α determines the tendency to gamble for large or small EVs whereas μ is analogous to an intercept parameter reflecting the tendency to gamble for a hypothetical EV of zero. Here, the context parameter τ is set to zero. (B) Effect of varying the value function parameter α and the context parameter τ is considered. Red lines represent agents with a positive value function coefficient α (equal to 0.15), and green lines represent agents with a negative alpha (equal to −0.15). Agents with different τ are plotted in which τ increases in £0.5 steps from −£2 to £2 along a bright-to-dark gradient. (C) Experiment 1: relationship between the effect of EV at current trial t and effect of EV at previous trial $t - 1$ on gambling probability ($r(24) = -.44$, $p = .033$). (D) Same analysis performed on data simulated with the computational model ($r(24) = -.605$, $p = .002$).

ensure that this did not bias our analyses, for each experiment we considered the winning model (see Results) and compared it with an equivalent model, except that the values of $\bar{r}_t$ and/or $\bar{r}_{t,k}$ at the start of the task were set as free parameters. For all experiments, these more complex models showed a larger BIC (Experiment 1: 13,310 vs. 13,381; Experiment 2: 14,330 vs. 14,415; fMRI experiment: 12,711 vs. 12,789), indicating that models with the values of $\bar{r}_t$ and/or $\bar{r}_{t,k}$ at the start of the task set as free parameters were overparameterized.

fMRI Scanning and Analysis

Details of the methods employed for the fMRI experiment have previously been reported (see also Rigoli, Rutledge, Dayan, et al., 2016). Visual stimuli were back-projected onto a translucent screen positioned behind the bore of the magnet and viewed via an angled mirror. BOLD contrast functional images were acquired with echo-planar T2*-weighted (EPI) imaging using a Siemens (Berlin, Germany) Trio 3-T MR system with a 32-channel head coil. To maximize the signal in our ROIs, a partial volume of the ventral part of the brain was recorded. Each image volume consisted of 25 interleaved 3-mm-thick sagittal slices (in-plane resolution = 3 × 3 mm, time to echo = 30 msec, repetition time = 1.75 sec). The first six volumes acquired were discarded to allow for T1 equilibration effects. T1-weighted structural images were acquired at a 1 × 1 × 1 mm resolution. fMRI data were analyzed using Statistical Parametric Mapping Version 8 (Wellcome Trust Centre for Neuroimaging). Data preprocessing included spatial realignment, unwarping using individual field maps, slice-timing correction, normalization, and smoothing. Specifically, functional volumes were realigned to the mean volume, were spatially normalized to the standard Montreal Neurological Institute template with a 3 × 3 × 3 voxel size, and were smoothed with 8-mm Gaussian kernel. Such kernel was used following previous studies from our lab, which used the same kernel to maximize the statistical power in midbrain regions (Rigoli, Chew, Dayan, & Dolan, 2016a; Rigoli, Friston, & Dolan, 2016; Rigoli, Rutledge, Dayan, et al., 2016). High-pass filtering with a cutoff of 128 sec and AR(1) model were applied.

For our analyses, neural activity was estimated with two general linear models (GLMs). Both GLMs were associated with a canonical hemodynamic function and included six nuisance motion regressors. The first GLM included a stick function regressor at option presentation modulated by a conventional RPE signal, corresponding to the actual EV of options minus the predicted EV of options. The predicted EV of options corresponds to the expected contextual reward $\bar{r}_t$ estimated with the computational model of choice behavior selected by model comparison (see below). This was estimated trial-by-trial using an equal learning rate $\eta = 0.51$ (i.e., the average within the sample) for all participants. The use of a single learning rate for all participants was motivated by considerations in favor of this approach compared with using participant-specific estimates in model-based fMRI (Wilson & Niv, 2015). To ascertain that our findings were not biased by the use of the same learning rate for all participants, we rerun the fMRI analyses below using individual learning rates and obtained the same findings (results not shown).

It has been pointed out that the separate components of the RPE (in our study, actual and predicted option EV) are correlated with the RPE, and so an area that is only reporting actual EV might falsely be seen as reporting a full RPE. Therefore, a better way to address our question (Niv et al., 2012; Niv & Schoenbaum, 2008) is to test for two findings: first, a negative correlation with the predicted EV and, second, a positive correlation with the actual EV. We followed this approach estimating a second GLM, which included a stick function regressor at option presentation modulated by two separate variables, one corresponding to the actual EV of options and the other to the predicted EV of options. These two parametric modulators were only mildly correlated (max Pearson coefficient across participants $r = .2$) and were included symmetrically in the GLM model, allowing us to estimate their impact on neural activation in an unbiased way.

The GLMs also included a stick function regressor at outcome presentation modulated by an outcome prediction error corresponding to the difference between the choice outcome and the actual EV of options. The outcome prediction error was equivalent to zero for choices of sure options and was either positive (for reward outcomes) or negative (for zero outcomes) for choices of gambles.

Note that, at the behavioral level, the predicted EV of options could potentially depend on a cue-dependent component (associated with explicit instructions) and a cue-independent component (derived from learning). Being our focus on learning, we aimed at isolating the contribution of the latter. To this aim, we exploited the fact that each block was associated with a single contextual cue (including 140 trials), implying that the cue-dependent component was constant within a block whereas the cue-independent component varied. We estimated the GLMs separately for each of the four blocks, a procedure which allowed us to isolate the contribution of the cue-independent component related with learning.

Contrasts of interest were computed participant by participant and used for second-level one-sample t tests and regressions across participants. Substantial literature motivated us to restrict statistical testing to a priori ROIs: VTA/SN and ventral striatum (Lak et al., 2014; Stauffer et al., 2014; Niv et al., 2012; Park et al., 2012; D'Ardenne et al., 2008; Tobler et al., 2005; O'Doherty et al., 2003, 2004; Schultz et al., 1997). For VTA/SN, we used bilateral anatomical masks manually defined using the software MRIcro and the mean structural image for the group, similar to the approach used in Guitart-Masip et al. (2011). For ventral striatum, we used an 8-mm sphere centered on coordinates from a recent meta-analysis on incentive value processing (left striatum: −12, 12, −6; right striatum: 12, 10, −6; Bartra et al., 2013). For hypothesis testing, we adopted voxel-wise small volume correction (SVC) with a $p < .05$ family-wise error as significance threshold.

RESULTS

Experiment 1

The goal of this experiment was to assess whether participants learn contextual reward expectation from previous experience and, if so, at what rate. The average gambling percentage did not differ from 50% across participants (mean = 46%, *SD* = 22%; $t(23) = -0.95, p = .35$). The lack of risk aversion is consistent with prior reports using a similar task (Rigoli, Friston, & Dolan, 2016; Rigoli, Friston, Martinelli, et al., 2016; Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016) and may reflect the use of small monetary payoffs (Prelec & Loewenstein, 1991). By design, the sure option and the gamble had always equivalent EV and the EV at trial t was uncorrelated with the EV at trial $t - 1$ ($t(23) = 1; p = .5$). For each participant, we built a logistic regression model of choice (i.e., with dependent measure being choice of the gamble or of the sure option) which included the EV at trial t and the EV at trial $t - 1$ as regressors. Across participants, the slope coefficient associated with EV at trial t did not differ from zero (mean = 0.11, *SD* = 0.95; $t(23) = 0.54, p = .59$), whereas the slope coefficient associated with EV at trial $t - 1$ was significantly less than zero (mean = −0.04, *SD* = 0.01; $t(23) = -2.28, p = .032$), indicating participants gambled more with smaller EVs at trial $t - 1$. To investigate whether choice was influenced more by the EV at trial t or by the EV at trial $t - 1$, we computed the absolute value of the slope parameter associated with the first and second variable in the logistic regression. The absolute value of the slope coefficient associated with EV at trial t was larger than the absolute value of the slope coefficient associated with EV at trial $t - 1$ ($t(23) = 3.62, p = .002$), indicating that the EV at trial t exerted a greater influence than the EV at trial $t - 1$ on choice.

Our main hypothesis was that contextual reward expectations are learned from previous trials and drive choice adaptation. Such learning implies that the expected contextual reward at trial t is lower/higher when a low/high EV is presented at trial $t - 1$. We derived our predictions about choice adaptation from previous data (Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016), which show that, consistent with adaptation to context, participants who prefer to gamble for large EVs (at trial t) gamble more when EVs are larger relative to contextual expectations, whereas participants who prefer to gamble for small EVs gamble more when EVs are smaller relative to contextual expectations. These considerations led us to predict a relationship between the effect of EV at trial t and the effect of EV at trial $t - 1$ on gambling percentage. Consistent with this prediction, we observed an inverse correlation across individuals between the slope coefficient associated with EV at trial t and the slope coefficient associated with EV at trial $t - 1$ ($r(24) = -.44, p = .033$; Figure 2C; this result is still significant when using a Kendall correlation, which is less affected by extreme values; $t(24) = -0.29, p = .047$). This indicates that participants who gambled more with larger EVs at trial t also gambled more with smaller EVs at trial $t - 1$, and participants who gambled more with smaller EVs at trial t also gambled more with larger EVs at trial $t - 1$.

To consider the influence of previous EVs further, we fitted to gambling data an exponential decay model (see Methods and Equation 1). Consistent with an influence exerted by previous trials, we found an inverse correlation between the weight parameters β_1 and β_2 (Supplementary Figure S4; $r(24) = -.52, p = .009$). The median decay parameter λ was equal to 1.54, which implies that a weight β_2 at trial $t - 1$ will become $\beta_2/4.5$ at trial $t - 2$. To compare the impact on choice of the EV at trial t against the overall impact of EVs at previous trials, we considered the absolute value of β_1 and of $\sum_{j=1}^{4} \beta_2 e^{-\lambda(j-1)}$ and found no difference between these two quantities ($t(23) = 1.08, p = .29$).

Next, we compared different generative models of choice behavior (see Methods). According to BIC scores (see Table 1), in the selected model (i) an average reward was learned from previous trials and exerted value adaptation, (ii) a constant (and not decreasing) learning rate was implemented, and (iii) normalization was subtractive (and not divisive). Consistent with adaptation to context, the context parameter τ of the selected model (which is multiplied by the average reward, and the total is subtracted to the EV) was significantly larger than zero (Supplementary Figure S1; $t(23) = 3.23, p = .004$). The median learning rate η of the selected model was 0.68.

We used the full model and participant-specific parameter estimates of that model to generate simulated choice behavioral data and perform behavioral analyses on the ensuing data. The model replicated the main statistical result from the raw data, namely the correlation between the effect on choice (i.e., the slope coefficient of logistic regression of choice) of EV at trial t and of EV at trial $t - 1$ ($r(24) = -.605, p = .002$; Figure 2D), an effect not replicated using a model with a decreasing learning rate ($r(24) = -.08, p = .726$).

Overall, these results show that reward expectations about options can be learned from recent experience and that subjective values are adapted to these expectations with an impact on choice behavior. In addition, data suggest that this form of learning is based on a constant learning rate (and not a quickly decaying learning rate) and that adaptation is subtractive (and not divisive).

Table 1. Model Comparison Analysis for the Three Experiments

Model	*Free Param*	*Neg LL*	*BIC*	*N Sub*
Experiment 1				
Random	–	9132	18264	0
Slope only	α	7314	14779	0
Intercept only	μ	7601	15353	3
Slope and intercept	μ, α	6509	13321	4
Subtractive; $\bar{r}_t$; constant η	μ, α, η, τ	6352	13310***	17
Subtractive; $\bar{r}_t$; decreasing η	μ, α, τ	6499	13452	0
Divisive; $\bar{r}_t$; constant η	μ, α, η, τ	6409	13424	0
Divisive; $\bar{r}_t$; decreasing η	μ, α, τ	6500	13454	0
Experiment 2				
Random	–	9283	18567	0
Slope only	α	7869	15895	0
Intercept only	μ	7903	15965	7
Slope and intercept	μ, α	7016	14345	3
Subtractive; $\bar{r}_t$; constant η	μ, α, η, τ	6990	14607	0
Subtractive; $\bar{r}_t$; decreasing η	μ, α, τ	6999	14468	0
Subtractive; $\bar{r}_{t,k}$; constant η	μ, α, η, τ	6919	14465	0
Subtractive; $\bar{r}_{t,k}$; decreasing η	μ, α, τ	6930	14330***	16
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η	μ, α, η, τ	6927	14482	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η	μ, α, τ	6938	14346	2
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	μ, α, η, τ	6908	14444	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	μ, α, η, τ	6910	14448	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η	$\mu, \alpha, \eta, \tau_1, \tau_2$	6924	14633	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η	$\mu, \alpha, \tau_1, \tau_2$	6915	14457	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	$\mu, \alpha, \eta, \tau_1, \tau_2$	6912	14609	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	$\mu, \alpha, \eta, \tau_1, \tau_2$	6915	14616	0
Divisive; $\bar{r}_t$; constant η	μ, α, η, τ	6988	14603	0
Divisive; $\bar{r}_t$; decreasing η	μ, α, τ	6994	14460	0
Divisive; $\bar{r}_{t,k}$; constant η	μ, α, η, τ	6940	14509	0
Divisive; $\bar{r}_{t,k}$; decreasing η	μ, α, τ	6936	14342	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η	μ, α, η, τ	6966	14559	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η	μ, α, τ	6951	14373	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	μ, α, η, τ	6929	14484	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	μ, α, η, τ	6950	14528	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η	$\mu, \alpha, \eta, \tau_1, \tau_2$	6952	14687	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η	$\mu, \alpha, \tau_1, \tau_2$	6900	14427	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	$\mu, \alpha, \eta, \tau_1, \tau_2$	6909	14603	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	$\mu, \alpha, \eta, \tau_1, \tau_2$	6911	14607	0

Table 1. (*continued*)

Model	*Free Param*	*Neg LL*	*BIC*	*N Sub*
fMRI Experiment				
Random	–	8122	16245	0
Slope only	α	7091	14316	0
Intercept only	μ	7033	14198	5
Slope and intercept	μ, α	6279	12824	3
Subtractive; $\bar{r}_t$; constant η	μ, α, η, τ	6130	12792	0
Subtractive; $\bar{r}_t$; decreasing η	μ, α, τ	6229	12857	0
Subtractive; $\bar{r}_{t,k}$; constant η	μ, α, η, τ	6127	12785	0
Subtractive; $\bar{r}_{t,k}$; decreasing η	μ, α, τ	6199	12796	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η	μ, α, η, τ	6105	12742	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η	μ, α, τ	6177	12753	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	μ, α, η, τ	6090	12711***	13
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	μ, α, η, τ	6099	12730	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η	μ, α, η, τ_1, τ_2	6099	12864	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η	μ, α, τ_1, τ_2	6116	12762	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	μ, α, η, τ_1, τ_2	6098	12861	0
Subtractive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	μ, α, η, τ_1, τ_2	6088	12841	0
Divisive; $\bar{r}_t$; constant η	μ, α, η, τ	6157	12846	0
Divisive; $\bar{r}_t$; decreasing η	μ, α, τ	6237	12872	0
Divisive; $\bar{r}_{t,k}$; constant η	μ, α, η, τ	6154	12840	0
Divisive; $\bar{r}_{t,k}$; decreasing η	μ, α, τ	6208	12815	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η	μ, α, η, τ	6155	12843	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η	μ, α, τ	6192	12782	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	μ, α, η, τ	6154	12839	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with single τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	μ, α, η, τ	6170	12871	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η	μ, α, η, τ_1, τ_2	6143	12950	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η	μ, α, τ_1, τ_2	6157	12846	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; constant η for $\bar{r}_t$; decreasing η for $\bar{r}_{t,k}$	μ, α, η, τ_1, τ_2	6140	12943	0
Divisive; $\bar{r}_t$, $\bar{r}_{t,k}$ with multiple τ; decreasing η for $\bar{r}_t$; constant η for $\bar{r}_{t,k}$	μ, α, η, τ_1, τ_2	6114	12893	0

For models considered (organized in rows), columns report (from left to right) (i) model description, indicating whether divisive or subtractive normalization is implemented, whether adaptation involves $\bar{r}_t$, $\bar{r}_{t,k}$, or both (and in the latter case whether a single or multiple context parameter τ is implemented), and whether learning involves a constant of decreasing learning rate η; (ii) negative log-likelihood (Neg LL), estimated separately for each individual's choice data (excluding trials with RTs slower than 3 sec and faster than 300 msec) and summed across subjects; (iii) free parameters (Free Param); (iv) BIC (models with the lowest BIC are marked with asterisks); and (v) number of subjects (N Sub) for which the model shows the lowest BIC.

Experiment 2

The second experiment assessed whether learning an average reward expectation takes account of an alternation of context, which is signaled by distinct cues. In principle, two forms of learning can be considered: (i) first, participants might learn the reward available at previous trials independent of any cue, similar to Experiment 1, and (ii) second, participants might differentiate between contexts and learn an average reward representation specific for each cue.

We first investigated learning by ignoring changes in cues. We analyzed the relationship between the EV at trial t and the EV at trial $t - 1$ as in Experiment 1. The average gambling percentage did not differ from 50% across participants (mean = 55, *SD* = 21; *t*(24) = 1.24, *p* = .23). By design, the EV of options at trial *t* was correlated weakly with the EV at trial *t* − 1 (max Pearson coefficient across participants *r* = .19). We built a logistic regression model of choice (having choice of the gamble or of the sure option as dependent measure), which included the EVs at trial *t* and trial *t* − 1 as regressors. The slope coefficient associated with EV at trial *t* was not significantly different from zero (mean = −0.02, *SD* = 0.54; *t*(24) = −0.21, *p* = .83), whereas the slope coefficient associated with EV at trial *t* − 1 was significantly smaller than zero (mean = −0.07, *SD* = 0.17; *t*(24) = −3.09, *p* = .005), indicating that participants overall gambled more with smaller EVs at trial *t* − 1. To investigate whether choice was influenced more by the EV at trial *t* or by the EV at trial *t* − 1, we computed the absolute value of the slopes associated with the EV at trial *t* and with the EV at trial *t* − 1 in the logistic regression. The absolute value of the slope coefficient associated with EV at trial *t* was larger than the absolute value of the slope coefficient associated with EV at trial *t* − 1 (*t*(24) = 3.57, *p* = .002), indicating that the EV at trial *t* exerted a greater influence than the EV at trial *t* − 1 on choice. We performed a similar analysis as for Experiment 1, which showed an inverse correlation between the slope coefficients associated with EV at trial *t* and the slope coefficients associated with EV at trial *t* − 1 (*r*(25) = −.46, *p* = .021; Figure 3A). This indicates that participants who gambled more with larger EVs at trial *t* also gambled more with smaller EVs at trial *t* − 1, and participants who gambled more with smaller EVs at trial *t* also gambled more with larger EVs at trial *t* − 1.

We next considered the hypothesis that the two alternating cues have an impact on learning and value adaptation independent of previous trials. To address this question, we analyzed the second half of the task when knowledge of context contingencies is likely to be more secure. Here, we focused only on the very first trial of each block, and among these trials, we considered those associated with £3 and £5 EV, as these are common to both the high- and low-value contexts. We predicted that, for these trials, participants would exhibit different preferences dependent on the context condition. Consistent with this prediction, we found a correlation between the effect on gambling of EV at trial *t* (i.e., the slope of a logistic regression model having EV at trial *t* as regressor) and the difference in gambling between low- and high-value contexts for EVs common to both contexts (*r*(25) = .46, *p* = .020). In other words, participants who overall gambled more with larger EVs also gambled more when the common EVs were relatively larger in the context of the new block, whereas participants who overall gambled more with smaller EVs also gambled more when common EVs were relatively smaller in the context associated with the new block. Here, the focus on first trials of

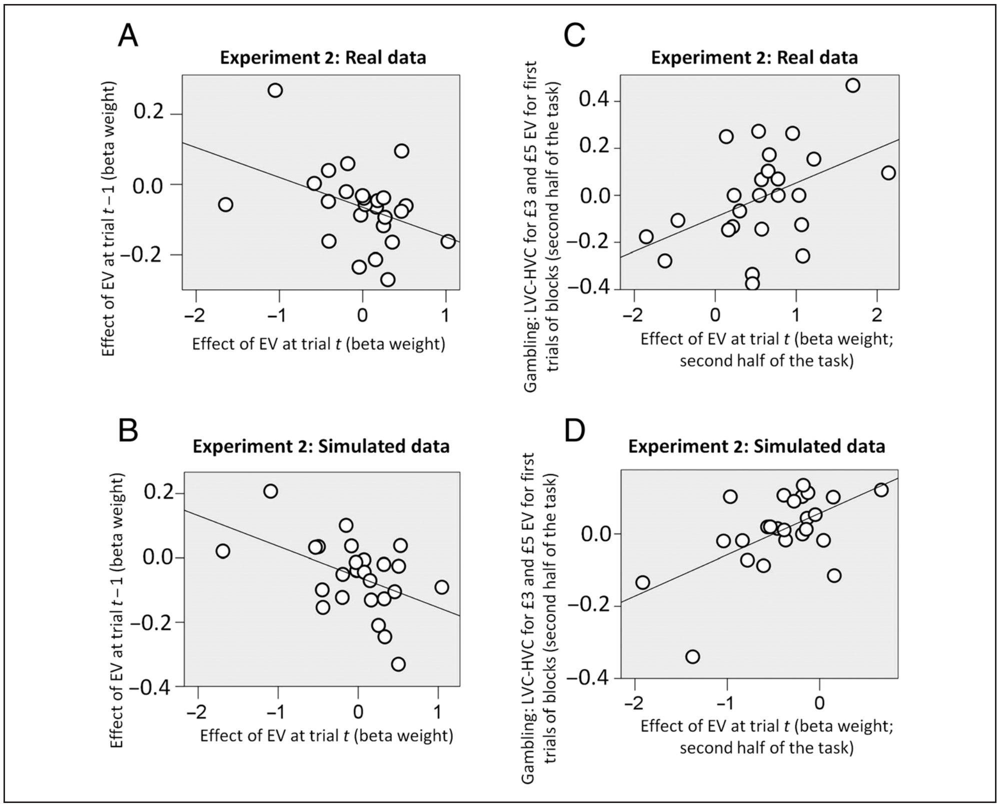


Figure 3. (A) Experiment 2: relationship between the effect of EV at current trial *t* and effect of EV at previous trial *t* − 1 on gambling probability (*r*(25) = −.46, *p* = .021). (B) Same analysis performed on data simulated with the computational model (*r*(25) = −.48, *p* = .016). (C) Experiment 2: analysis of effect of context. Relationship between the effect on gambling of EV at trial *t* (i.e., the slope of a logistic regression model having EV at trial *t* as regressor) and the difference in gambling between low- and high-value contexts for EVs common to both contexts (associated with £3 and £5 EV; *r*(25) = .46, *p* = .020), only considering first trials of blocks and the second half of the task. Since the slope of the logistic regression is estimated from the second half of the task only, note that it is different from the one estimated from the whole task (shown in A). (D) Same analysis performed on data simulated with the computational model (*r*(25) = .49, *p* = .01).

blocks is crucial, because for these trials, the EVs presented previously are orthogonal to the current context condition, allowing us to show a context effect independent of previous trials.

To probe further the mechanisms underlying learning and context sensitivity, we compared different generative models of choice behavior (see Methods). According to BIC scores (see Table 1), in the selected model (i) a cue-dependent average reward was learned and exerted value adaptation, whereas a cue-independent average reward was not implemented; (ii) a decreasing (and not constant) learning rate characterized learning; and (iii) normalization was subtractive (and not divisive). In the selected model, the context parameter τ is multiplied by a cue-dependent reward representation (in turn acquired following a decreasing learning rate), and the total is subtracted to the option EV. Consistent with adaptation to context, the context parameter τ of the selected model was significantly larger than zero (Supplementary Figure S2; $t(24) = 2.11, p = .045$).

We used the selected model and participant-specific parameter estimates from that model to generate simulated choice behavioral data and perform behavioral analyses on the ensuing data. The selected model replicated the correlation between the effect on choice (i.e., the slope coefficient of logistic regression model of choice) of EV at trial t and of EV at trial $t - 1$ ($r(25) = -.48, p = .016$; Figure 3B), an effect not replicated with a model without the context parameter τ ($r(25) = -.17, p = .42$). The selected model also replicated the correlation between the effect on choice of EV at trial t and the difference in gambling for low- minus high-value context for EVs common to both contexts, when considering first trials of blocks (and focusing on the second half of the task; $r(25) = .58, p = .002$; Figure 3D). This correlation was not replicated with a model implementing an average reward independent of context and a constant learning rate ($r(25) = .17, p = .41$). These results indicate that, when multiple contexts alternate, value and choice adaptation can be driven by a representation of the two context averages, without learning based on previous reward experience independent of cues. In addition, data suggest that learning of contextual reward representations is based on a decreasing learning rate and that adaptation is subtractive.

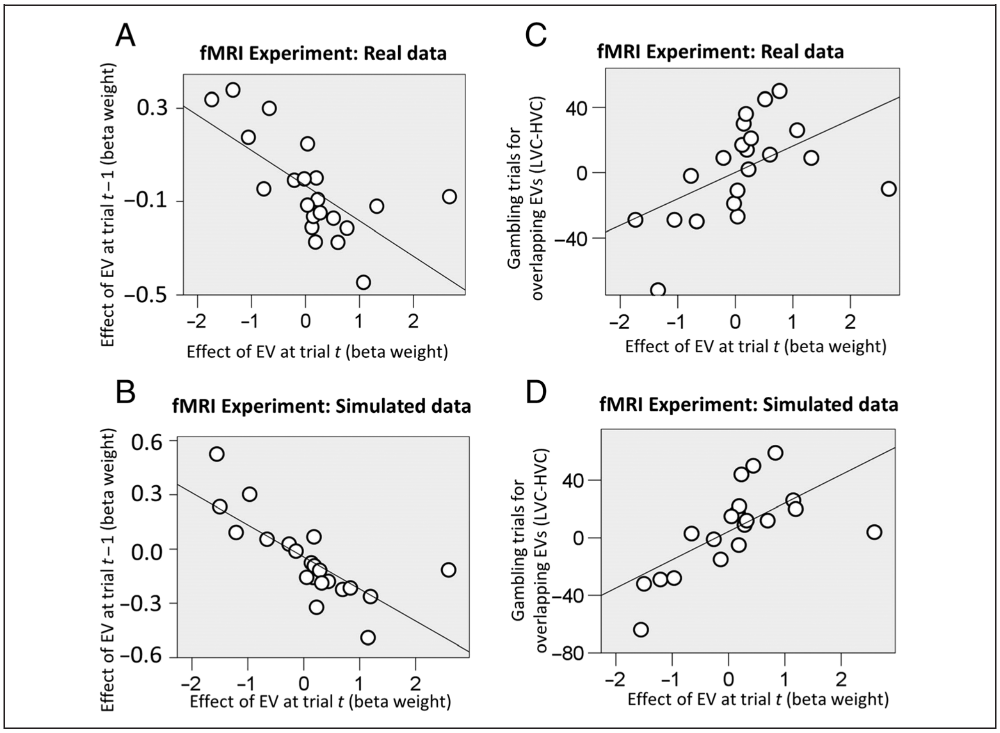
fMRI Experiment

The results of both experiments motivated us to reanalyze data from an fMRI experiment involving a similar task (Rigoli, Rutledge, Dayan, et al., 2016). The paradigm was similar to the task used in Experiment 2, since both comprise two different contexts characterized by distinct reward distributions. However, the fMRI blocks were longer (around 10 min rather than the 30 sec of Experiment 2). We asked whether the presence of longer blocks is uninfluential on the contextual learning processes involved or whether it implies the recruitment of different processes. Critically, the characteristics of the context were presented to the participants explicitly before the start—so learning would formally have been unnecessary. In addition, the use of simultaneous fMRI recording allowed us to study the neural substrates of learning average reward representations.

The average gambling percentage did not differ from 50% across participants (mean = 51.5, $SD = 21.27$; $t(20) = 0.32, p = .75$). By design, the EV at trial t was only mildly correlated with the EV at trial $t - 1$ (max Pearson coefficient across participants $r = .13$). We built a logistic regression model of choice (having choice of the gamble or of the sure option as dependent measure), which included the EV at trial t and the EV at trial $t - 1$ as regressors. The slope coefficient associated with EV at trial t did not differ from zero (mean = 0.19, $SD = 1.07$; $t(20) = 0.81, p = .43$), nor did the slope coefficient associated with EV at trial $t - 1$ (mean = -0.05, $SD = 0.20$; $t(20) = -1.128, p = .27$). To investigate whether choice was influenced more by the EV at trial t or by the EV at trial $t - 1$, we computed the absolute value of the slopes associated with the EV at trial t and with the EV at trial $t - 1$ in the logistic regression. The absolute value of the slope coefficient associated with EV at trial t was larger than the absolute value of the slope coefficient associated with EV at trial $t - 1$ ($t(20) = 3.63$, $p = .002$), indicating that the EV at trial t exerted a greater influence than the EV at trial $t - 1$ on choice.

As for the previous experiments, we analyzed the effects of previous EVs ignoring cues and found an inverse correlation between the slope coefficient associated with EV at trial t and the slope coefficient associated with EV at trial $t - 1$ ($r(21) = -.64, p = .002$; Figure 4C). This indicates that participants who gambled more with larger EVs at trial t also gambled more with smaller EVs at trial $t - 1$, and participants who gambled more with smaller EVs at trial t also gambled more with larger EVs at trial $t - 1$. To consider the influence of previous EVs further, we fitted to gambling data an exponential decay model (see Methods; Equation 1). Consistent with an influence of previous trials, we found an inverse correlation between the weight parameters β_1 (linked with the influence of EV at trial t) and β_2 (linked with the exponentially decaying influence of EV at trials before t; Supplementary Figure S4; $r(21) = -.82, p < .001$). The median decay parameter λ was equal to 0.31, which implies that a weight β_2 at trial $t - 1$ will become $\beta_2/1.4$ at trial $t - 2$. We compared the decay parameter λ found here with the decay parameter λ found in Experiment 1, and the former was significantly smaller than the latter ($t(43) = 2.66, p = .011$), indicating that previous trials beyond $t - 1$ exerted a greater impact in the fMRI experiment compared with Experiment 1. To compare the impact on choice of the EV at trial t against the overall impact of EVs at previous trials, we considered the absolute value of β_1 and of $\sum_{j=1}^{4} \beta_2 e^{-\lambda(j-1)}$ and found no difference between the two quantities ($t(20) = 0.64, p = .53$).

Figure 4. (A) fMRI experiment: relationship between the effect of EV at current trial t and effect of EV at previous trial $t - 1$ on gambling probability ($r(21) = -.65$, $p < .002$). (B) Same analysis performed on data simulated with the computational model ($r(21) = -.75, p < .001$). (C) fMRI experiment: relationship between the effect of EV at current trial t and the number of gambling trials when comparing low-value context (LVC) and high-value context (HVC) for EVs common to both context ($r(21) = .56, p = .008$). (D) Same analysis performed on data simulated with the computational model ($r(21) = .66, p < .001$).



In our previous study (Rigoli, Rutledge, Dayan, et al., 2016), we assessed whether the two cues exert an influence on choice consistent with value adaptation. For each participant, we computed the gambling proportion with EVs common to both contexts (i.e., associated with the £2–£5 range) for the low- minus high-value context and found that this difference correlated with the effect of EV on gambling (as estimated with the logistic regression above; $r(21) = .56, p = .008$; Figure 4A). This is consistent with the idea that the two cues were considered during value computation and choice, though it is also compatible with an influence of previous reward experience independent of cues. Because the fMRI experiment involved four blocks alone, the task did not allow us to isolate effects on the very first trial of each block, as we did for Experiment 2, an analysis that could potentially have provided evidence of an independent role of cues.

To clarify further the relative impact of cue-dependent and cue-independent learning, we compared different generative models of choice behavior (see Methods). According to BIC scores (see Table 1), in the selected model (i) both a cue-independent and a cue-dependent average reward were learned and exerted value adaptation, (ii) a constant (and not decreasing) learning rate characterized learning of an average reward independent of cue, (iii) a decreasing (and not constant) learning rate characterized learning of an average reward associated with contextual cues, (iv) normalization was subtractive (and not divisive), and (v) a single context parameter was implemented for both a cue-independent and a cue-dependent average reward. Consistent with adaptation to context, the context parameter τ of the selected model was significantly larger than zero (Supplementary Figure S3; $t(20) = 4.02, p < .001$). The median learning rate η of the selected model was 0.37 ($\eta > 0.1$ for 16 participants). Notably, the model selected in the fMRI experiment is different from the model selected in Experiment 2; possible reasons explaining why this difference was observed are discussed below.

We used the selected model and participant-specific parameter estimates from that model to generate simulated choice behavioral data and perform behavioral analyses on the ensuing data. The selected model replicated the correlation between the gambling proportion with EVs common to both contexts (i.e., associated with £3 and £5) for the low- minus high-value context and the effect of EV on gambling (as estimated with the logistic regression above; $r(21) = .66, p < .001$; Figure 4B). This correlation was not replicated when using a model without the context parameter τ ($r(21) = -.14, p = .53$). The full model also replicated the correlation between the effect on choice (i.e., the slope coefficient of logistic regression model of choice) of EV at trial t and of EV at trial $t - 1$ ($r(21) = -.75, p < .001$; Figure 4D), an effect not replicated with a model without the context parameter τ ($r(21) = -.05, p = .82$).

Overall, we found that cue-dependent and cue-independent forms of learning could coexist with both affecting value and choice adaptation. These mapped into two distinct learning processes, with cue-dependent learning driven by a decreasing learning rate and cue-independent learning mediated via a constant learning rate. In addition, cue-dependent and cue-independent average rewards appeared to exert equal effects on value adaptation, which, as in previous experiments, was subtractive rather than divisive.

Finally, we reanalyzed fMRI data acquired during task performance. It is well established that, at outcome delivery, a response in ventral striatum and VTA/SN correlates positively and negatively with the actual and predicted reward, respectively, whereas in the same regions, at option presentation, a correlation with actual option EV is reported (Lak et al., 2014; Stauffer et al., 2014; Niv et al., 2012; Park et al., 2012; D'Ardenne et al., 2008; Tobler et al., 2005; O'Doherty et al., 2003, 2004; Schultz et al., 1997). These findings motivated a proposal of a distinct role of these regions at outcome delivery and option presentation, corresponding to signaling RPE and EV, respectively (Bartra et al., 2013). However, other theoretical models (Schultz et al., 1997) imply that dopaminergic regions reflect RPE also at option presentation. The difference between the two hypotheses is that the latter (Schultz et al., 1997), but not the former (Bartra et al., 2013), predicts that at option presentation neural activity inversely correlates with the predicted option EV, corresponding to the contextual average reward. However, this prediction has never been formally tested, and here we provide such test.

Neural response was first modeled using a GLM that included, at option presentation, a stick function regressor modulated by the actual EV of options minus the predicted EV of options (the latter corresponds to the average reward $\bar{r}_t$ learned from previous trials as prescribed by the computational model of choice behavior—see Methods). This parametric modulator, which represents a conventional RPE signal, correlated with activation in VTA/SN (3, −13, −14; $Z = 3.16, p = .032$ SVC) and ventral striatum (left: −12, 11, −2; $Z = 3.99, p = .002$ SVC; right: 9, 11, −2; $Z = 4.48, p < .001$ SVC).

Next, as a more stringent test, neural response was modeled using a second GLM, which included, at option presentation, a stick function regressor associated with two separate parametric modulators, one for the actual EV of options and the other for the predicted EV of options. A correlation with actual EV of options (Figure 5A–B) was observed in VTA/SN (9, −13, −17; $Z = 3.25, p = .028$ SVC) and ventral striatum (left: −12, 8, −2; $Z = 3.73, p = .005$ SVC; right: 9, 8, −2; $Z = 4.25, p = .001$ SVC), together with an inverse correlation with the average reward $\bar{r}_t$ (Figure 6A–B; VTA/SN: 12, −19, −11; $Z = 3.26, p = .011$ SVC; left ventral striatum: −12, 8, 1; $Z = 3.14, p = .026$ SVC; right ventral striatum: 18, 14, −2; $Z = 2.98, p = .039$ SVC). These results are consistent with an encoding of RPE signal after option presentation.

Encoding of a context-related RPE in VTA/SN and ventral striatum may represent a neural substrate mediating choice adaptation to context. If this was the case, we would predict a stronger neural sensitivity to contextual reward expectations in participants showing an increased influence of context on choice behavior (captured by the context parameter τ in our behavioral model). Consistent with this prediction, we observed an inverse correlation between the effect of predicted EV of options on neural response and the individual context parameter τ (estimated with the selected model of choice behavior) in VTA/SN (Figure 6C; 6, −19, −8; $Z = 2.90, p = .027$ SVC), but not in ventral striatum.

Overall, these findings indicate that activity in VTA/SN and ventral striatum increases with actual EV of options and decreases with the EV of options predicted based on recent trials, consistent with reflecting an RPE signal relative to average reward representations. In addition, response adaptation in VTA/SN (but not in ventral striatum) was linked with contextual adaptation in choice behavior.

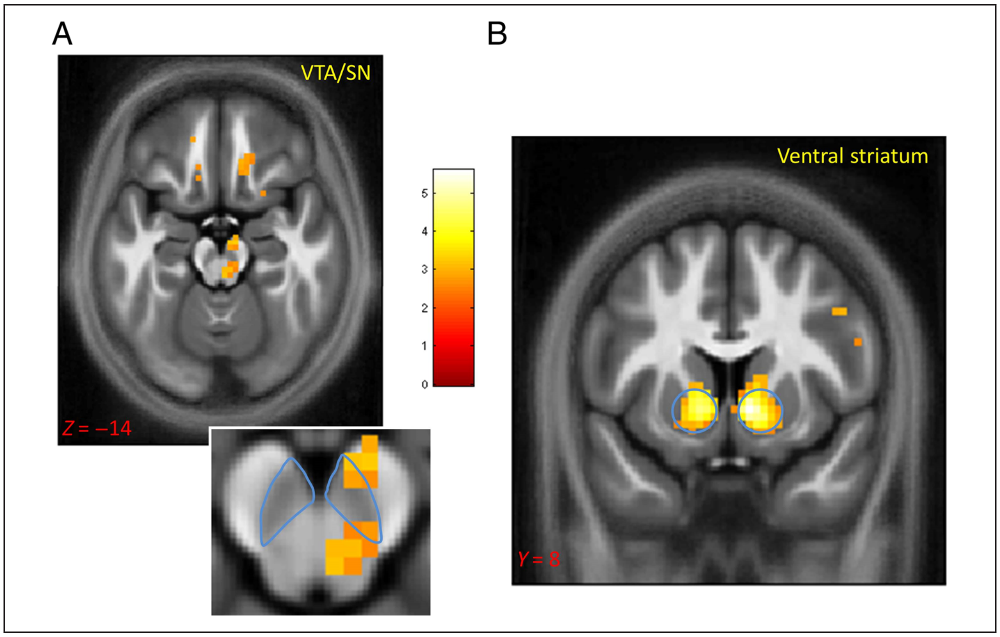


Figure 5. Activity at option presentation in our ROIs for a positive correlation with the actual EV of options. For display purposes, we show activity for voxels where the statistic is significant when using $p < .005$ uncorrected. (A) Activity shown for VTA/SN (9, −13, −17; $Z = 3.25$, $p = .028$ SVC; Montreal Neurological Institute coordinate space is used). (B) Activity shown for ventral striatum (left: −12, 8, −2; $Z = 3.73, p = .005$ SVC; right: 9, 8, −2; $Z = 4.25$, $p = .001$ SVC).

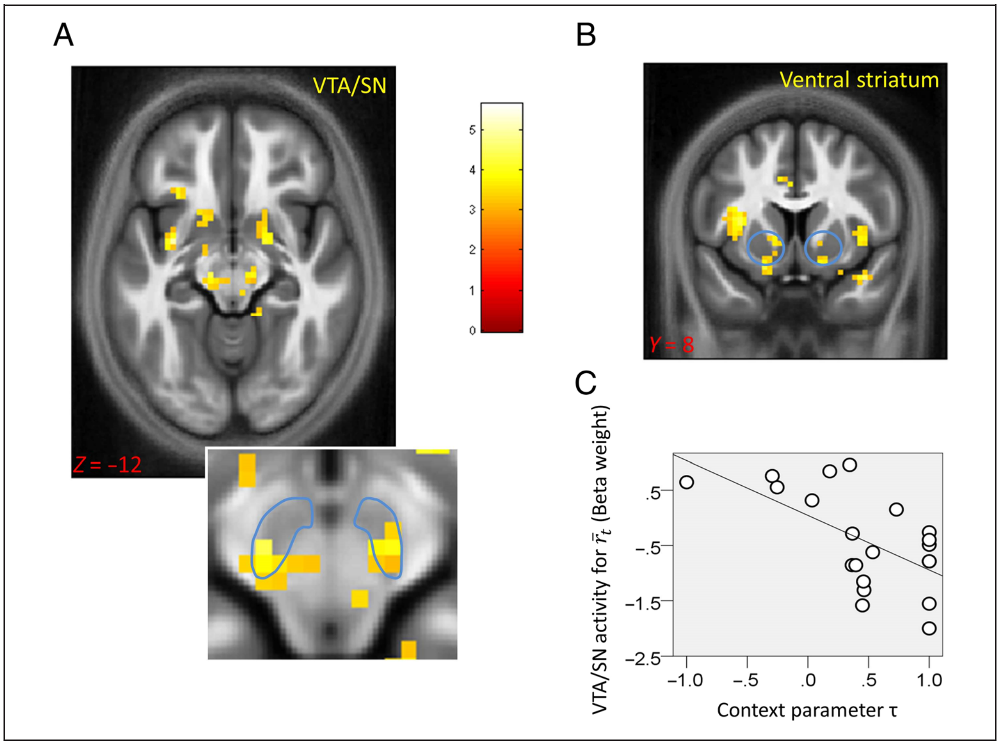


Figure 6. Activity at option presentation in our ROIs for a negative correlation with the predicted EV of options (estimated with the computational model of choice behavior, corresponding to the expected contextual reward). For display purposes, we show activity for voxels where the statistic is significant when using $p < .005$ uncorrected. (A) Activity shown for VTA/SN (VTA/SN: 12, −19, −11; Z = 3.26, p = .011 SVC). (B) Activity shown for ventral striatum (left ventral striatum: −12, 8, 1; Z = 3.14, p = .026 SVC; right ventral striatum: 18, 14, −2; Z = 2.98, p = .039 SVC). (C) Relationship between the behavioral context parameter τ (estimated with the computational model for each participant and indicating the degree of choice adaptation to the average reward learned from previous trials independent of context) and the beta weight for the correlation between VTA/SN activity and expected contextual reward (6, −19, −8; Z = 2.90, p = .027 SVC).

DISCUSSION

Contextual effects on choice depend on adaptation of incentive values to the average reward expected before option presentation (Rigoli, Friston, & Dolan, 2016; Rigoli, Friston, Martinelli, et al., 2016; Rigoli, Rutledge, Chew, et al., 2016; Rigoli, Rutledge, Dayan, et al., 2016; Louie et al., 2013, 2014, 2015; Summerfield & Tsetsos, 2015; Cheadle et al., 2014; Ludvig et al., 2014; Summerfield & Tsetsos, 2012; Carandini & Heeger, 2011; Stewart, 2009; Stewart, Chater, & Brown, 2006; Stewart et al., 2003). However, as explicit information about context was provided in previous studies, how contextual reward expectation is learned through experience remains poorly understood. Our study builds upon previous research on how the brain learns distributions of variables (Diederen, Spencer, Vestergaard, Fletcher, & Schultz, 2016; Nassar et al., 2012; Berniker, Voss, & Kording, 2010; Nassar, Wilson, Heasly, & Gold, 2010; Behrens, Woolrich, Walton, & Rushworth, 2007). However, as far as we are aware, none of the existing tasks have considered discrete choices (rather than estimation). Thus, we used a task in which a contextual distribution is learned from experience and adaptation to that distribution is expressed via discrete choices. We show that experience can drive learning of contextual reward expectations that in turn impact on value adaptation. This form of learning can be characterized using a model where, after an option is presented, the belief about an average reward is updated according to an RPE (i.e., the actual minus the predicted option EV) multiplied by a learning rate. The average reward expectation acquired through learning in turn elicits subtractive (and not divisive) normalization by setting a reference point to which option values are rescaled, influencing choice behavior.

In Experiment 1, option EVs were drawn from a single reward distribution. Consistent with some models (Niv et al., 2007), participants learned an average reward representation from previous trials, which was updated following a constant learning rate (Rescorla & Wagner, 1972). However, contrary to predictions from these models (Niv et al., 2007), we observed a large learning rate implying that recent (and not long-run) experience is relevant. Data from Experiment 2, where two contexts characterized by distinct reward distributions alternated at a fast rate, showed no evidence of cue-independent learning. Instead, they highlight a cue-dependent learning whereby different reward representations were acquired in association with contextual cues. This form of learning was characterized by a decreasing learning rate, implying that experience early in the task is weighted more than later experience. This can be formally described with Bayesian learning assuming fixed reward statistics of the context

(Bishop, 2006) and is linked to previous associative learning theories (Dayan, Kakade, & Montague, 2000; Pearce & Hall, 1980).

A reanalysis of a previous data set (Rigoli, Rutledge, Dayan, et al., 2016) shows cue-independent learning based on recent past reward experience (similar to Experiment 1) combined with learning based on contextual cues (similar to Experiment 2). Here, value and choice adaptation were affected by reward representations arising from both forms of learning. Why the coexistence of these two learning components emerged here, but not in Experiment 2, remains to be fully understood. One important difference between the two tasks is in block length, with Experiment 2 having short (30-sec) blocks and the fMRI experiment long (10-min) blocks. Furthermore, explicit information regarding contextual reward distribution was provided in the fMRI experiment, entailing that participants did not need to learn the distribution. One possibility is that learning from recent reward experience and attending to fast-changing contextual cues (as in Experiment 2) are demanding cognitive processes that compete against each other, leading to reliance on the latter process alone (which is more informative about upcoming reward). By contrast, attending to slow-changing contextual cues or knowing explicitly the contextual reward distributions (as in the fMRI experiment) might make fewer demands on cognitive resources, allowing participants to attend to both contextual cues and past reward experience. Investigating this hypothesis requires an assessment of the relative amount of cognitive resources necessary to attend fast- and slow-changing contextual cues, respectively.

An important question arising from our findings is on the link between the learning mechanisms identified here and cognitive functions such working memory. A possibility is that cue-dependent learning recruits working memory, at least when contexts alternate rapidly. This is supported by our observation that cue-dependent learning suppresses cue-independent learning when cues alternate rapidly, suggesting the involvement of high-demanding cognitive functions including working memory. Under such conditions, fast and flexible working memory mechanisms would be most useful. Moreover, the observation of a decreasing learning rate characterizing cue-dependent learning may be consistent with the involvement of working memory, whereby during the initial stage beliefs update quickly and are retrieved flexibly thereafter. Although these aspects support the involvement of working memory during cue-dependent learning (at least when cues alternate quickly), we note that another fundamental feature of working memory is limited capacity, which implies a loss of information if the cognitive load increases. We did not manipulate the cognitive load (as, for instance, did Collins & Frank, 2012). Research that assesses the impact of cognitive load on the learning mechanism studied here would be necessary to establish the involvement of working memory.

The coexistence (at least in some conditions) of cue-dependent and cue-independent learning (with an impact on value adaptation) leads to questions about their relationship. One possibility is that a unique brain system is responsible for computing both representations. Alternatively, different brain systems may be involved. For instance, the VTA/SN may mediate learning from recent reward experience independent of any cue-related information, whereas hippocampus may mediate cue-dependent learning (Rigoli, Friston, & Dolan, 2016; Wimmer & Shohamy, 2012; Rudy, 2009; Shohamy, Myers, Hopkins, Sage, & Gluck, 2009; Doeller, King, & Burgess, 2008; Fanselow, 2000; Holland & Bouton, 1999). This possibility is indirectly supported by our findings that cue-independent learning is guided by a constant learning rate whereas cue-dependent learning is driven by a decaying learning rate. There is a parallel between the difference in learning rate found here and differences in neural processing observed in VTA/SN, striatum, and amygdala on the one hand and the hippocampus on the other (Rudy, 2009; Marschner, Kalisch, Vervliet, Vansteenwegen, & Büchel, 2008; Matus-Amat, Higgins, Barrientos, & Rudy, 2004; Fanselow, 2000; Holland & Bouton, 1999). Though our task does not imply any hierarchical order between the two forms of learning that emerged here, one possibility is that they map to different hierarchical levels in the participants' model of the world, as described formally by hierarchical Dirichlet process and hierarchical reinforcement learning models (Botvinick, Niv, & Barto, 2009; Barto & Mahadevan, 2003).

Previous research has left open the question of whether presenting options elicits a response in VTA/SN and ventral striatum that reflects the average EV of options (Bartra et al., 2013) or an RPE signal (Schultz et al., 1997). One important previous study did analyze RPE signaling at the time of option presentation (Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008). However, in that study, at the time of option presentation participants received also a monetary outcome that was independent of choice, and that outcome was included in the analysis as a component of the RPE signal. In other words, in the study of Hare et al. (2008), the effects of outcome and option are combined, as they occur simultaneously and are analyzed together. We sought to examine an unconfounded case in which there are only options and no outcome. We address this question showing that, consistent with an RPE signal dependent on presenting options, activity in ventral striatum and VTA/SN was characterized by a positive and negative correlation with actual and predicted option EV, respectively. These results indicate that activity in the striatum and VTA/SN reflects predictions about option EV that correspond to the contextual reward. This indicates that neural representations of EV predictions are not fixed but evolve on the basis of previous experience. In addition, these findings show that the temporal dynamics of this form of learning in the brain reflect the dynamics observed in choice behavior, as both indicate that EV predictions depend on recent—and not long-run—experience.

The activity of dopaminergic neurons in VTA/SN and the release of this neuromodulator in the ventral striatum play a central role in signaling RPE during learning with single rewards (Lak et al., 2014; Stauffer et al., 2014; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Tobler et al., 2005; Schultz et al., 1997). An influential model proposes that phasic dopamine responses encode RPE signals whereas tonic dopamine activity reflects beliefs about average reward (Niv et al., 2007). Our data support the idea that dopaminergic regions process information about average reward, though they highlight a phasic (i.e., RPE signaling), rather than tonic, response associated with average reward (see also Diuk, Tsai, Wallis, Botvinick, & Niv, 2013). Although fMRI does not allow us to make neurochemical inferences, one possibility is that the context-related RPE signal found here is mediated by some aspect of dopaminergic functioning. We emphasize also that our analyses are uninformative regarding the role of tonic neural activity (e.g., linked with dopamine). Further research is necessary to elucidate the role of the latter in contextual reward representations formed during choice behavior, though links have been reported between tonic activity in dopaminergic regions and representations of average reward in other domains (Hamid et al., 2016; Rigoli et al., 2016a). Influential views propose a motivational role for dopamine and average reward in energizing behavior (Niv et al., 2007; Salamone & Correa, 2002; Dickinson, Smith, & Mirenowicz, 2000; Berridge & Robinson, 1998). The link between motivational vigor and average reward in the context of choice is potentially interesting but remains to be investigated.

In VTA/SN (but not in ventral striatum), the degree of correlation between average reward and brain activity was associated with choice adaptation to context. In other words, for participants whose choice behavior was affected more by expectations about option EVs, VTA/SN response was also affected more by reward expectations, consistent with the possibility that signaling in VTA/SN might mediate learning and value adaptation as expressed in behavior. The finding of a correlation between adaptation in VTA/SN and adaptation in choice is consistent with previous reports (Rigoli, Friston, & Dolan, 2016; Rigoli, Rutledge, Dayan, et al., 2016). Here we extend on these previous findings by showing that the relationship between VTA/SN and behavioral adaptation emerges also when average reward expectation is learned from previous trials.

The current data suggest various directions for future studies. It is promising to take advantage of the richer picture of contexts provided by forms of nonparametric Bayesian generative modeling (Collins & Frank, 2013; Gershman, Blei, & Niv, 2010; Gershman & Niv, 2010; Redish, Jensen, Johnson, & Kurth-Nelson, 2007; Courville, Daw, & Touretzky, 2006; Daw, Courville, & Touretzky, 2006), possibly hierarchically, whereby participants can generate their own notion of context. Another direction is inspired by evidence that, in addition to adapting to the mean of rewards, response in many brain regions adapts to reward variability (Cox & Kable, 2014; Park et al., 2012; Bermudez & Schultz, 2010; Kobayashi et al., 2010; Rorie et al., 2010; Padoa-Schioppa, 2009; Padoa-Schioppa & Assad, 2008; Tobler et al., 2005). An open question is whether adaptation to variability characterizes subjective value and choice and, if so, how representations of reward variability are learned. A third direction is to examine the intricate complexities of temporal adaptation apparent in sensory systems (Panzeri, Brunel, Logothetis, & Kayser, 2010; Wark, Fairhall, & Rieke, 2009; Kording, Tenenbaum, & Shadmehr, 2007; Fairhall, Lewen, Bialek, & van Steveninck, 2001) or the second order effects of alternating volatilities (Behrens et al., 2007). A fourth direction would be to consider avoidance of punishments as well as the acquisition of rewards (Rigoli, Chew, Dayan, & Dolan, 2016b; Rigoli, Pezzulo, & Dolan, 2016; Rigoli, Pavone, & Pezzulo, 2012).

In summary, we show that experience drives learning of contextual reward expectations to which subjective values are adapted. Learning supports the acquisition of both cue-related and cue-unrelated reward expectations. We clarify the neural substrates of learning contextual reward representations highlighting an encoding of context-related RPE in VTA/SN and ventral striatum, with activity in the former region linked with choice adaptation to context. These findings are relevant for understanding the connection between reward learning and context sensitivity.

Acknowledgments

This work was supported by the Wellcome Trust (Ray Dolan Senior Investigator Award 098362/Z/12/Z) and the Max Planck Society. P. D. was supported by the Gatsby Charitable Foundation. The Wellcome Trust Centre for Neuroimaging was supported by core funding from the Wellcome Trust 091593/Z/10/Z. We would like to thank Robb Rutledge and Cristina Martinelli for helpful discussions on the topic of the study.

Reprint requests should be sent to Francesco Rigoli, The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, 12 Queen Square, London, WC1N 3BG, UK, or via e-mail: f.rigoli@ucl.ac.uk.

REFERENCES

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems, 13,* 341–379.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage, 76,* 412–427.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10,* 1214–1221.

Bermudez, M. A., & Schultz, W. (2010). Reward magnitude coding in primate amygdala neurons. *Journal of Neurophysiology, 104,* 3424–3432.

Berniker, M., Voss, M., & Kording, K. (2010). Learning priors for Bayesian computations in the nervous system. *PLoS One, 5,* e12686.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28,* 309–369.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

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

Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13,* 51–62.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., De Gardelle, V., Castañón, S. H., et al. (2014). Adaptive gain control during human perceptual choice. *Neuron, 81,* 1429–1441.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience, 35,* 1024–1035.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review, 120,* 190–229.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10,* 294–300.

Cox, K. M., & Kable, J. W. (2014). BOLD subjective value signals exhibit robust range adaptation. *Journal of Neuroscience, 34,* 16533–16543.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science, 319,* 1264–1267.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII, 23,* 3–38.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation, 18,* 1637–1677.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience, 3,* 1218–1223.

Dickinson, A., Smith, J., & Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behavioral Neuroscience, 114,* 468–483.

Diederen, K. M., Spencer, T., Vestergaard, M. D., Fletcher, P. C., & Schultz, W. (2016). Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency. *Neuron, 90,* 1127–1138.

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

Doeller, C. F., King, J. A., & Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences, U.S.A., 105,* 5915–5920.

Fairhall, A. L., Lewen, G. D., Bialek, W., & van Steveninck, R. R. D. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature, 412,* 787–792.

Fanselow, M. S. (2000). Contextual fear, gestalt memories, and the hippocampus. *Behavioural Brain Research, 110,* 73–81.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117,* 197–209.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology, 20,* 251–256.

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J., Dayan, P., Dolan, R. J., et al. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *Journal of Neuroscience, 31,* 7867–7875.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., et al. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience, 19,* 117–126.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience, 28,* 5623–5630.

Holland, P. C., & Bouton, M. E. (1999). Hippocampus and context in classical conditioning. *Current Opinion in Neurobiology, 9,* 195–202.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9,* 90–98.

Kobayashi, S., de Carvalho, O. P., & Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *Journal of Neuroscience, 30,* 534–544.

Kording, K. P., Tenenbaum, J. B., & Shadmehr, R. (2007). The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature Neuroscience, 10,* 779–786.

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review, 112,* 841–861.

Lak, A., Stauffer, W. R., & Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proceedings of the National Academy of Sciences, U.S.A., 111,* 2343–2348.

Louie, K., Glimcher, P. W., & Webb, R. (2015). Adaptive neural coding: From biological to behavioral decision-making. *Current Opinion in Behavioral Sciences, 5,* 91–99.

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences, U.S.A., 110,* 6139–6144.

Louie, K., LoFaro, T., Webb, R., & Glimcher, P. W. (2014). Dynamic divisive normalization predicts time-varying value coding in decision-related circuits. *Journal of Neuroscience, 34,* 16046–16057.

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making, 27,* 146–156.

Marschner, A., Kalisch, R., Vervliet, B., Vansteenwegen, D., & Büchel, C. (2008). Dissociable roles for the hippocampus and the amygdala in human cued versus context fear conditioning. *Journal of Neuroscience, 28,* 9030–9036.

Matus-Amat, P., Higgins, E. A., Barrientos, R. M., & Rudy, J. W. (2004). The role of the dorsal hippocampus in the acquisition and retrieval of context memory representations. *Journal of Neuroscience, 24,* 2431–2439.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience, 15,* 1040–1046.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience, 30,* 12366–12378.

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology, 191,* 507–520.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience, 32,* 551–562.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences, 12,* 265–272.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38,* 329–337.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304,* 452–454.

Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience, 29,* 14004–14014.

Padoa-Schioppa, C., & Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nature Neuroscience, 11,* 95–102.

Panzeri, S., Brunel, N., Logothetis, N. K., & Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. *Trends in Neurosciences, 33,* 111–120.

Park, S. Q., Kahnt, T., Talmi, D., Rieskamp, J., Dolan, R. J., & Heekeren, H. R. (2012). Adaptive coding of reward prediction errors is gated by striatal coupling. *Proceedings of the National Academy of Sciences, U.S.A., 109,* 4285–4289.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532–552.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature, 442,* 1042–1045.

Prelec, D., & Loewenstein, G. (1991). Decision making over time and under uncertainty: A common approach. *Management Science, 37,* 770–786.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16,* 323–341.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review, 114,* 784–805.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory, 2,* 64–99.

Rigoli, F., Chew, B., Dayan, P., & Dolan, R. J. (2016a). The dopaminergic midbrain mediates an effect of average reward on Pavlovian vigor. *Journal of Cognitive Neuroscience, 28,* 1303–1317.

Rigoli, F., Chew, B., Dayan, P., & Dolan, R. J. (2016b). Multiple value signals in dopaminergic midbrain and their role in avoidance contexts. *Neuroimage, 135,* 197–203.

Rigoli, F., Friston, K. J., & Dolan, R. J. (2016). Neural processes mediating contextual influences on human choice behaviour. *Nature Communications, 7,* 12416.

Rigoli, F., Friston, K. J., Martinelli, C., Selaković, M., Shergill, S. S., & Dolan, R. J. (2016). A Bayesian model of context-sensitive value attribution. *eLife, 5,* e16127.

Rigoli, F., Pavone, E. F., & Pezzulo, G. (2012). Aversive Pavlovian responses affect human instrumental motor performance. *Frontiers in Neuroscience, 6,* 134.

Rigoli, F., Pezzulo, G., & Dolan, R. J. (2016). Prospective and Pavlovian mechanisms in aversive behaviour. *Cognition, 146,* 415–425.

Rigoli, F., Rutledge, R. B., Chew, B., Ousdal, O. T., Dayan, P., & Dolan, R. J. (2016). Dopamine increases a value-independent gambling propensity. *Neuropsychopharmacology, 41,* 2658–2667.

Rigoli, F., Rutledge, R. B., Dayan, P., & Dolan, R. J. (2016). The influence of contextual reward statistics on risk preference. *Neuroimage, 128,* 74–84.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108,* 370–392.

Rorie, A. E., Gao, J., McClelland, J. L., & Newsome, W. T. (2010). Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey. *PLoS One, 5,* e9308.

Rudy, J. W. (2009). Context representations, context functions, and the parahippocampal–hippocampal system. *Learning & Memory, 16,* 573–585.

Salamone, J. D., & Correa, M. (2002). Motivational views of reinforcement: Implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behavioural Brain Research, 137,* 3–25.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.

Shohamy, D., Myers, C. E., Hopkins, R. O., Sage, J., & Gluck, M. A. (2009). Distinct hippocampal and basal ganglia contributions to probabilistic learning and reversal. *Journal of Cognitive Neuroscience, 21,* 1821–1833.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research, 29,* 281–295.

Soltani, A., De Martino, B., & Camerer, C. (2012). A range-normalization model of context-dependent choice: A new model and evidence. *PLoS Computational Biology, 8,* e1002607.

Stauffer, W. R., Lak, A., & Schultz, W. (2014). Dopamine reward prediction error responses reflect marginal utility. *Current Biology, 24,* 2491–2500.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology, 62,* 1041–1062.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1–26.

Stewart, N., Chater, N., Stott, H. P., & Reimers, S. (2003). Prospect relativity: How choice options influence decision under risk. *Journal of Experimental Psychology: General, 132,* 23–46.

Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience, 6,* 70.

Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences, 19,* 27–34.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science, 307,* 1642–1645.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences, U.S.A., 109,* 9659–9664.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences, U.S.A., 113,* 3102–3107.

Tsetsos, K., Usher, M., & Chater, N. (2010). Preference reversal in multiattribute choice. *Psychological Review, 117,* 1275–1293.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79,* 281–299.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review, 111,* 757–769.

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. (2011). Does the brain calculate value? *Trends in Cognitive Sciences, 15,* 546–554.

Wark, B., Fairhall, A., & Rieke, F. (2009). Timescales of inference in visual adaptation. *Neuron, 61,* 750–761.

Wilson, R. C., & Niv, Y. (2015). Is model fitting necessary for model-based fMRI? *PLoS Computational Biology, 11,* e1004237.

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