frontiers | Frontiers in Behavioral Neuroscience

TYPE Original Research
PUBLISHED 09 November 2022
DOI 10.3389/fnbeh.2022.1041566

OPEN ACCESS

EDITED BY
Alberto Del Arco,
University of Mississippi, United States

REVIEWED BY
Ron Keiflin,
University of California, Santa Barbara, United States
Baihan Lin,
Columbia University, United States

*CORRESPONDENCE
Kristoffer C. Aberg
kc.aberg@gmail.com

SPECIALTY SECTION
This article was submitted to Motivation and Reward, a section of the journal Frontiers in Behavioral Neuroscience

RECEIVED 11 September 2022
ACCEPTED 26 October 2022
PUBLISHED 09 November 2022

CITATION
Aberg KC and Paz R (2022) Average reward rates enable motivational transfer across independent reinforcement learning tasks. *Front. Behav. Neurosci.* 16:1041566. doi: 10.3389/fnbeh.2022.1041566

Average reward rates enable motivational transfer across independent reinforcement learning tasks

Kristoffer C. Aberg* and Rony Paz

Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel

Outcomes and feedbacks on performance may influence behavior beyond the context in which it was received, yet it remains unclear what neurobehavioral mechanisms may account for such lingering influences on behavior. The average reward rate (ARR) has been suggested to regulate motivated behavior, and was found to interact with dopamine-sensitive cognitive processes, such as vigilance and associative memory encoding. The ARR could therefore provide a bridge between independent tasks when these are performed in temporal proximity, such that the reward rate obtained in one task could influence performance in a second subsequent task. Reinforcement learning depends on the coding of prediction error signals by dopamine neurons and their downstream targets, in particular the nucleus accumbens. Because these brain regions also respond to changes in ARR, reinforcement learning may be vulnerable to changes in ARR. To test this hypothesis, we designed a novel paradigm in which participants (n = 245) performed two probabilistic reinforcement learning tasks presented in interleaved trials. The ARR was controlled by an "induction" task which provided feedback with a low (p = 0.58), a medium (p = 0.75), or a high probability of reward (p = 0.92), while the impact of ARR on reinforcement learning was tested by a second "reference" task with a constant reward probability (p = 0.75). We find that performance was significantly lower in the reference task when the induction task provided low reward probabilities (i.e., during low levels of ARR), as compared to the medium and high ARR conditions. Behavioral modeling further revealed that the influence of ARR is best described by models which accumulates average rewards (rather than average prediction errors), and where the ARR directly modulates the prediction error signal (rather than affecting learning rates or exploration). Our results demonstrate how affective information in one domain may transfer and affect motivated behavior in other domains. These findings are particularly relevant for understanding mood disorders, but may also inform abnormal behaviors attributed to dopamine dysfunction.

KEYWORDS

reinforcement learning, average reward, motivation, prediction error, behavioral modeling, reward, transfer, controllability

Introduction

Negative information, such as bad news or negative feedback, may affect behavior in subsequent everyday-life tasks that are unrelated to the received information. Despite being relevant for understanding the impact of, for example, mood-related dysfunctions on cognitive performance, few studies investigated the neurocomputational correlates of how information received in one task may linger and affect behavior in other unrelated tasks.

One theory posits that motivated behavior depends on the average reward rate (ARR), such that vigilance and response rates increase in contexts were rewards are frequently provided (Niv, 2007; Niv et al., 2007). In support, human research report that response vigor (Guitart-Masip et al., 2011; Rigoli et al., 2016), as well as memory encoding success (Aberg et al., 2017, 2020), are affected by ARRs. Accordingly, the ARR may provide a bridge which enables recently received affective information (e.g., negative news or feedbacks) to temporally extend and influence subsequent motivational states and task performances.

The motivational impact of average reward levels was coupled with BOLD signal in a midbrain region pertaining to the loci of dopamine neurons (Aberg et al., 2020), while other studies report associations between ARRs and dopamine release in the nucleus accumbens (Hamid et al., 2016; Mohebi et al., 2019). Reinforcement learning depends on the neural representation of prediction errors (i.e., the mismatch between an actual and a predicted outcome; Sutton and Barto, 2018), which are coded by midbrain dopamine neurons and their downstream targets, including the nucleus accumbens (Rutledge et al., 2010; Hart et al., 2014; Schultz, 2016). Accordingly, reinforcement learning tasks may be particularly vulnerable to manipulations of average reward levels.

Moreover, average reward levels may affect learning performance *via* different computational mechanisms, e.g., *via* altered learning rates, decision biases, or reward processing, and this may be due to an accumulation of prediction errors (Eldar and Niv, 2015) or rewards (Aberg et al., 2020). To disentangle the impact of average reward on these different computational mechanisms, we designed and confronted a number of different behavioral models.

Here, participants performed two reinforcement learning tasks presented in interleaved trials. An "inducer" task was used to control the ARR by providing probabilistic rewards with a low, medium, or a high probability. A "reference" task, with a constant (medium) reward probability, was used to estimate the impact of ARRs on learning performance in the different conditions. In three separate experiments, we observed lower learning performance in the reference task when it was presented together with an inducer task that provided probabilistic rewards with a low probability (as compared to medium and high probabilities). Furthermore, careful behavioral modeling revealed that the impact of ARRs was best described by models which (i) accumulate average rewards (rather than prediction errors), and (ii) allow the ARR to directly modulate the prediction error signal (rather than learning rates or decision making biases).

Abnormal reinforcement learning patterns may play a role in the acquisition and maintenance of dysfunctional behaviors in relation to psychiatric and neurological disorders (Maia and Frank, 2011). Therefore, understanding how reinforcement learning is affected by transfer of affective information between tasks, and how these related to motivated behaviors and interact with the dopamine system, is relevant for psychopathology.

Materials and methods

Participants

The study followed the declaration of Helsinki and was approved by the Institutional Review Board (IRB) of the Weizmann Institute. Informed consent was provided before the start of the testing. In total, 245 participants were recruited *via* Amazon Mechanical Turk (n experiment 1/2/3 = 63/79/103). Inclusion criteria consisted of being older than 18 years, speaking English fluently, and having completed more than 95% of previously started assignments on the Amazon Mechanical Turk platform. Each participant could only perform one of the three experiments, and were recruited *via* identical ads on the Amazon Mechanical Turk platform, and their participation was decided on a first-come-first-served bases. Participants were excluded from the data analysis if their average overall performance on the last four trials was less than 0.6, if they failed to respond on more than 20 trials, or if they exhibited more than two response sequences where the same button was pressed more than ten times in a row. After applying these exclusion criteria, data from 148 participants were included in the analyses (n experiment 1/2/3 = 50/58/40).

Task and procedure

Participants performed probabilistic reinforcement learning tasks, where in each trial one object in a pair of objects was selected (**Figure 1A**). Probabilistic feedback was then presented based on the reward/punishment probability assigned to the selected object. For example, as illustrated in **Figure 1A**, selecting the apple would yield positive feedback with probability p_{Reward}, while selecting the pear would yield negative feedback with probability $p_{Punishment}$. In the present study, p_{Reward} was always equal to $p_{Punishment}$. Critically, to control the ARR, here defined as average rewards per trials, an "inducer" task was performed in interleaved trials together with a "reference" task. The "inducer" task provided positive feedback with p_{Reward} set to either high (11/12,

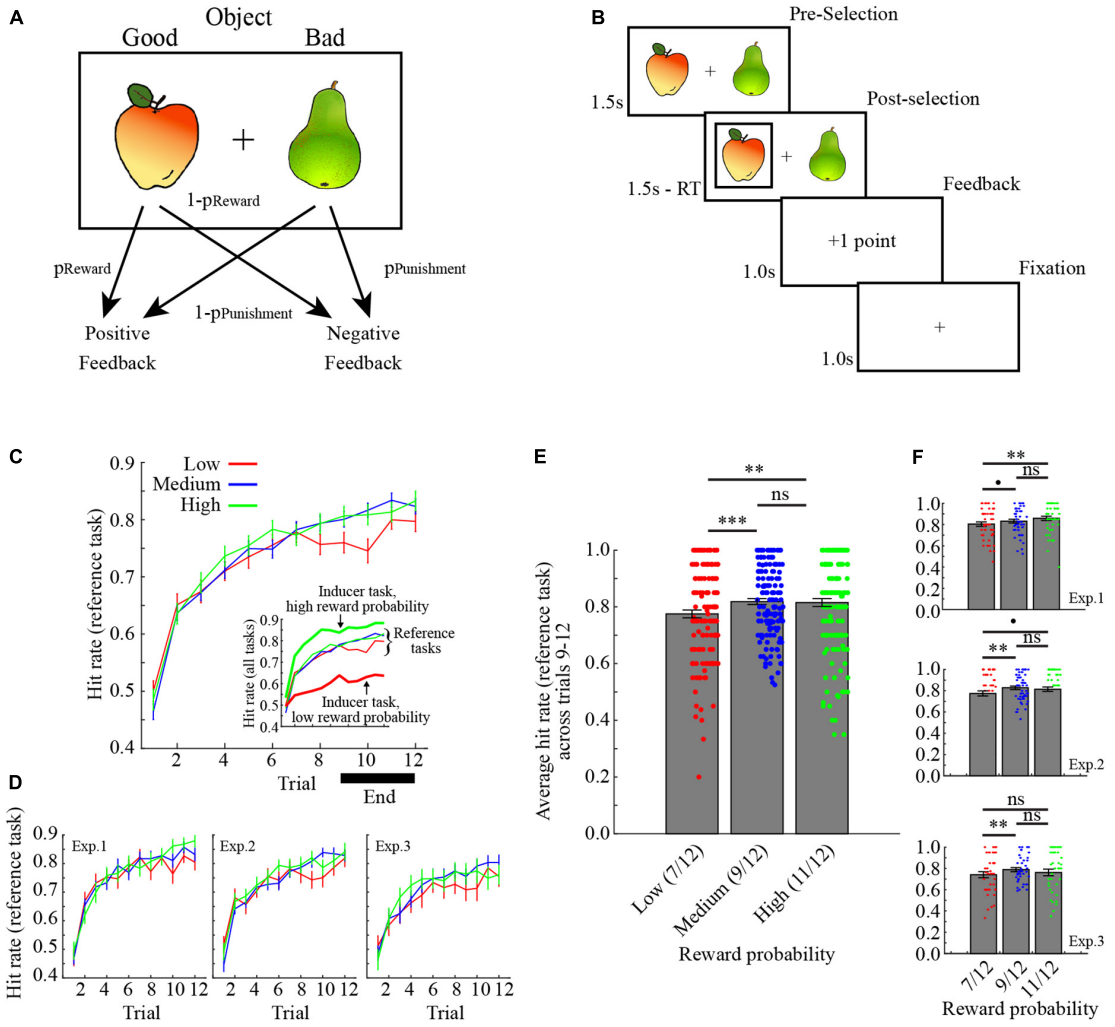

FIGURE 1
Task description and learning performance. **(A)** In each trial, participants selected one of two options in a pair. One object provided positive feedback with probability p_{Reward}, while the other object provided negative feedback with probability $p_{Punishment}$. Here, p_{Reward} was always equal to $p_{Punishment}$. **(B)** Illustration of a trial. **(C)** Learning curves for the reference tasks, collapsed across the three experiments. The inset also shows performance for the inducer tasks with low- and high reward probabilities. Inducer/reference tasks are, respectively denoted by bold / narrow lines. **(D)** Learning curves for the reference task in the three experiments separately. **(E)** Average learning performance for the reference tasks across trials 9–12 collapsed across the three experiments. **(F)** Average learning performance for the reference tasks across trials 9–12 separately for the three experiments. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $p < 0.10$, ns $p > 0.10$. All error bars indicate the standard error of the mean.

p_{Reward}~0.92), medium (9/12, p_{Reward} = 0.75), or low (7/12, p_{Reward}~0.58), while for the "reference" task p_{Reward} was always set to medium (9/12, p_{Reward} = 0.75). In other words, the ARR was, respectively highest and lowest in conditions where the inducer task provided a high/low probability of reward.

Each trial started with a display of two images for a maximum of 1.5 s (Figure 1B). After selection, the selected object was marked for the remainder of the 1.5 s, followed by a feedback displayed for 1 s. A fixation cross was then displayed for 1 s before the start of the next trial. If participants failed to respond within 1.5 s, a screen displaying the sentence "Too Slow! −2 points" was displayed for 3 s.

Each task was performed for 12 trials in each block (for a total of 24 trials per block, i.e., 12 inducer task trials +12 reference task trials), and the order of trials for each task within a block was interleaved in a pseudorandom fashion, such that no task was repeated for more than three trials in a row (e.g., . . . Inducer-Reference-Reference-Inducer-Inducer . . . is correct, while . . . Inducer-Reference-Reference-Reference-Reference . . . is incorrect).

Each participant performed five blocks of trials for each of the three conditions (i.e., five blocks each with p_{Reward} of the inducer task set to high, medium, and low), for a total of 15 blocks and 360 trials. The order of blocks was pseudorandomized with the limitation that one block of each

condition had been performed before repeating a condition (e.g., ... Low-High-Medium, Low-Medium-High, Medium-High-Low ... is a correct example of three possible sequential blocks, while ... Low-Low-Medium, High-Low-Medium, High-High-Medium ... is an example of an incorrect sequence of three blocks). In each task, a new pair of objects were presented for a total of 32 different pairs of objects (15 blocks of two interleaved tasks in the experiment proper +1 training block).

Before the start of the experiment, participants received explicit instructions on how to perform the task. For example, in experiment 1 the instructions were:

> *"Two objects will be presented in each trial, one to the left and one to the right. Select the object on the left by pressing the 'LeftArrow' key and the object on the right by pressing the 'RightArrow' key. Selecting an object results in one of the following types of feedback:*
>
> *+1: you gained 1 point.*
>
> *−1: you lost 1 point.*
>
> *Importantly! Collecting points will earn you a monetary bonus.*
>
> *One object in each pair is more likely to give better feedback than the other.*
>
> *Increase your bonus by learning which are the best objects.*
>
> *Each pair of objects is presented for 1.5 s. Beware! If your response is too slow, you will lose 2 points.*
>
> *Press the SPACE bar to start the task."*

After having read the instructions, participants performed one training block before continuing to the experiment proper. The experiments lasted less than 30 min, and participants were paid $2 for their participations and up to $2 extra as a performance-based bonus.

Experiments

Three different experiments were conducted, all based on the procedure described above, but with different feedback values.

Experiment 1

Positive/negative feedbacks were, respectively +1/−1 points for both the inducer tasks and the reference task.

Experiment 2

Positive/negative feedbacks were set to +1/−1 points for the reference task, while positive/negative feedback was +0.6/−0.6 for the inducer task with high p_{Reward}, +1/−1 for the inducer task with medium p_{Reward}, and +3.0/−3.0 for the inducer task with low p_{Reward}. These specific feedback values provided the same average number of points if the best option is selected in all trials (i.e., low p_{Reward}: $7 \times 3/12 + 5 \times -3/12 = 0.5$ points/trial; medium p_{Reward}: $9 \times 1/12 + 3 \times -1/12 = 0.5$ points/trial; high p_{Reward}: $11 \times 0.6/12 + 1 \times -0.6/12 = 0.5$ points/trial).

Experiment 3

Positive/negative feedbacks were set to +1/0 points for both the inducer tasks and the reference task. This experiment is identical to experiment 1, except that the negative (−1 point) feedback was replaced by a neutral (0 point) feedback.

Behavioral modeling

All models are based on the Q-learning algorithm (Watkins and Dayan, 1992), where the expected value $Q_i(t)$ of the selected option i in trial t is updated by the mismatch between the expected value and the actual outcome $R(t)$, i.e., the prediction error $\delta_Q(t)$, scaled by the learning rate α_Q.

$$Q_i(t+1) = Q_i(t) + \alpha * \delta_Q(t) \quad (1)$$

$$\delta_Q(t) = R(t) - Q_i(t) \quad (2)$$

For all models, the probability p of selecting option i in trial t is modeled using a soft-max choice probability function:

$$p_i(t) = \frac{e^{Q_i(t)*\beta_Q}}{\sum_i e^{Q_i(t)*\beta_Q}} \quad (3)$$

The decision weight β_Q determines how strongly a decision is affected by expected values, such that small values of β_Q allows for more stochastic decisions/exploration (Gershman, 2018; Wilson and Collins, 2019).

A popular derivation of the Q-learning algorithm separates the learning rates for feedbacks that are better or worse than expected (i.e., positive and negative prediction errors):

$$Q_i(t+1) = Q_i(t) + \alpha_{\delta_{Q+}} * \delta_Q(t), \; if \; \delta_Q(t) > 0 \quad (4)$$

$$Q_i(t+1) = Q_i(t) + \alpha_{\delta_{Q-}} * \delta_Q(t), \; if \; \delta_Q(t) < 0 \quad (5)$$

While these two models provide good fits to behavior in similar tasks (Frank et al., 2009; Aberg et al., 2015), neither model allows two interleaved tasks to interact (i.e., performance on one task is independent from the feedbacks received in other tasks). For this reason, we created a new set of models which

allows the ARR μ to, respectively modulate the learning rate α, the decision weight β_Q, and the prediction error signal δ_Q:

$$\alpha\,(t) \;=\; \alpha_0 * (1+\mu\,(t)) \tag{6}$$

$$\beta_Q\,(t) \;=\; \beta_0 * (1+\mu\,(t)) \tag{7}$$

$$\delta_Q\,(t) \;=\; R(t) - Q_i\,(t) + \mu(t) \tag{8}$$

The manipulations in Equations 6, 7, and 8, respectively allows larger values of μ to increase the learning rates, reduce exploration, and boost the prediction error signal.

Furthermore, the average reward may affect behavior in two separate ways. First, *via* the average accumulation of prediction errors (Eldar and Niv, 2015):

$$\mu_{PE}\,(t+1) \;=\; \mu_{PE}\,(t) + \alpha_\mu * (\delta_Q\,(t) - \mu_{PE}\,(t)) \tag{9}$$

Second, *via* the average accumulation of rewards (Aberg et al., 2020):

$$\mu_R\,(t+1) \;=\; \mu_R\,(t) + \alpha_\mu * (R\,(t) - \mu_R\,(t)) \tag{10}$$

To determine which of these two mechanisms that provide the best fit to behavior in our study, the two different ways of estimating μ, i.e., μ_{PE} and μ_R, were included in different models.

Finally, based on the suggestion of one reviewer, we also included a model which tracks positive and negative outcomes in separate streams (Lin et al., 2020). Specifically, expected negative ($Q_{negative}$) and positive ($Q_{positive}$) outcomes are, respectively updated following negative ($R_{negative}$) and positive ($R_{positive}$) feedbacks:

$$Q_{i,positive}\,(t+1) \;=\; \lambda_+ * Q_{i,\ positive}\,(t) + \alpha * \delta_{Q,\ positive}(t) \tag{11}$$

$$\delta_{Q,\ positive}\,(t) \;=\; w_+ * R_{positive}(t) - Q_{i,\ positive} \tag{12}$$

$$Q_{i,negative}\,(t+1) \;=\; \lambda_- * Q_{i,\ negative}\,(t) + \alpha * \delta_{Q,\ negative}(t) \tag{13}$$

$$\delta_{Q,\ negative}\,(t) \;=\; w_- * R_{negative}(t) - Q_{i,\ negative} \tag{14}$$

This model allows the subjective weighting of negative and positive feedbacks *via* the free parameters w_- and w_+, as well as the discounting of previous positive and negative outcomes *via* the free parameters. λ_+ and λ_-. Decisions are made by considering the combination of positive and negative expected outcomes:

$$Q_{i,total} \;=\; Q_{i,positive} + Q_{i,negative} \tag{15}$$

Of note, some modifications of the original model were necessary to enable its fit to behavior in the present task. First, the original model presumes that both positive and negative feedbacks are presented in each trial, while in the present study only one feedback-type was presented. For this reason, when positive or negative feedbacks were received, $R_{negative}$ or $R_{positive}$ was, respectively set to 0. Second, in the original model the option with the largest total expected outcome [i.e., argmax(Q_{total})] is deterministically selected, an unrealistic assumption for human behavior in probabilistic reinforcement learning tasks. For this reason, decisions were modeled using the softmax probability function (Equation 3), with Q_{total} as the input. Importantly, this option still allows deterministic decision making (as would be indicated by fitting a very large β_Q).

In total, nine different models were fitted to behavior:

- The canonical "ØØ-α" model, which is made up of Equations 1–3 with two free parameters: A learning rate α and a decision weight β_Q.
- The "ØØ-α$PE_{+,-}$" model, which is made up of Equations 2–5 with three free parameters: Two separate learning rates for negative and positive prediction errors α_{δ_Q-}, α_{δ_Q+}, and a decision weight β_Q.
- The "ØØ-α$R_{+,-}$" model, which is made up of Equations 3 and 11–14 with six free parameters: A learning rateα, two separate discount factors for negative and positive expected outcomes λ_-, λ_+, two separate weights for negative and positive outcomes w_-, w_+, and a decision weight β_Q.
- The "FB-PE" model, which is made up of Equations 1, 3, 8, and 9 with three free parameters: A learning rate α, a learning rate for average reward α_μ, and a decision weight β_Q.
- The "α-PE" model, which is made up of Equations 1–3, 6, and 9 with three free parameters: A constant learning rate term (for when the average reward is zero) α_0, a learning rate for average reward α_μ, and a decision weight β_Q.
- The "β-PE" model, which is made up of Equations 1–3, 7, and 9 with three free parameters: A learning rateα, a constant decision weight (for when the average reward is zero) β_0, and a learning rate for average reward α_μ.
- The "FB-R" model, which is made up of Equations 1,3, 8, and 10 with three free parameters: A learning rate α, a learning rate for average reward α_μ, and a decision weight β_Q.
- The "α-R" model, which is made up of Equations 1–3, 6, and 10 with three free parameters: A constant learning rate term (for when the average reward is zero) α_0, a learning rate for average reward α_μ, and a decision weight β_Q.
- The "β-R" model, which is made up of Equations 1–3, 7, and 10 with three free parameters: A learning rateα, a constant decision weight (for when the average reward is zero) β_0, and a learning rate for average reward α_μ.

In summary, the "ØØ-α," "ØØ-α$PE_{+,-}$," and the "ØØ-α$R_{+,-}$" models presume no impact of ARRs on performance. By contrast, the "FB-PE," "α-PE," and "β-PE" models allow the ARR

to affect performance *via* an accumulation of prediction errors across tasks, while the "FB-R," "α-R," and "β-R" models allow the ARR to affect performance *via* an accumulation of feedbacks. Further, "FB-x," "α-x," and "β-x" models, respectively presume that the ARR affects performance by influencing the prediction error signal, the learning rate, and the decision weight.

All models were fitted to behavior and confronted using a hierarchical Bayesian inference (HBI) method (Piray et al., 2019). The HBI concurrently fits the free parameters and compares the considered models (while also correcting for differences in model complexity), something which allows constraining individual fits to group-level hierarchical priors. Additionally, the random effects approach used by the HBI calculates both group-level statistics and model evidence based on the posterior probability that the model explains each subject's choice data. The HBI method provides fitted model parameters for each subject, as well as protected exceedance probabilities for each set of compared models. The exceedance probability estimates the probability that a model is the most frequent model to explain the observed behaviors, as compared to all other considered models (Rigoux et al., 2014). The *protected* exceedance probability (PXP) is more conservative, by taking into account the possibility that none of the compared models is supported by the data. To further demonstrate the robustness of the model selection procedure, we also report the model frequency, which is how often each model was determined to be the "best" model across participants.

To ensure that the parameters included in the selected model are meaningful, the values of the parameters used to simulate behaviors need to be successfully recovered when re-fitting the model to these simulated behaviors (Wilson and Collins, 2019). To confirm that this is the case, we randomly selected values of model parameters within the range of the fitted values obtained from the selected model, and generated the behavior of 1,000 virtual participants. Next, the selected model was re-fitted to the generated behaviors, and correlation coefficients were calculated between the generating and the recovered parameters. For a parameter to be meaningful, these correlations should be significantly positive.

Results

Behavioral results

Low average reward rate in the induction task reduces learning performance in the independent reference task

Learning curves collapsed across experiments are shown in Figure 1C, and for each experiment individually in Figure 1D. Learning performance for the reference task in each condition was defined as the average hit rate across the last third of the trials (i.e., trials 9–12). The average learning performance for the reference task in the different conditions collapsed across experiments are shown in Figure 1E, and for each experiment separately in Figure 1F.

To assess the impact of the ARR manipulation on learning performance in the reference tasks, and potential differences between experiments, the data from all participants were added to a mixed-factor ANOVA with between-subject Experiment (experiment 1–3) and Condition (inducer task with low, medium, and high reward probability). The results indicate a significant main effect of Condition [F (2,290) = 8.96, $p < 0.001$, η_p^2 0.058, ANOVA], but no main effect of Experiment [F (2,145) = 2.62, p = 0.077, η_p^2 0.035, ANOVA], nor interaction between Experiment and Condition [F (4,290) = 0.60, p = 0.665, η_p^2 0.008, ANOVA]. For a full ANOVA table, see Table 1. The main effect of condition was due to significantly lower learning performance for the reference task in the low average reward condition, as compared to both the medium and the high average reward conditions [Figure 1E; low versus medium: t (147) = −4.063, 95% CI = −0.062, −0.022, $p < 0.001$, Cohen's d = 0.334, two-tailed t-test; low versus high: t (147) = −3.236, 95% CI = −0.063, −0.015, p = 0.002, Cohen's d = 0.266, two-tailed t-test]. By contrast, there was no difference in learning performance for the reference task in the medium and high average reward conditions [t (147) = 0.273, 95% CI = −0.017, 0.022, p = 0.785, Cohen's d = 0.023, two-tailed t-test]. Notably, learning performance for the reference task in the low average reward condition was consistently lower in each experiment separately (for ANOVAs and t-tests, see Table 2).

In summary, supporting our predictions of an interaction between the inducer task and the reference task, we observed significantly lowered learning performance when the reference task was paired with an inducer task which provided low average reward. Furthermore, because these effects were observed in each of the three experiments, it cannot be attributed to differences in average reward magnitudes (tested in experiment 2) or increased salience attributed to negatively valued feedbacks (tested in experiment 3). Put simply, our data

TABLE 1 Mixed-effect ANOVA for the average hit rate as a function of Experiment (experiment 1, 2, and 3) and Condition (low, medium, and high reward probability).

	Sum of squares	*df*	Mean square	*F*	*P*-value	η_p^2
(Intercept)	274.92	1	274.92	5491.1	<0.001	
Experiment	0.262	2	0.131	2.615	0.077	0.035
Error	7.260	145	0.050			
Condition	0.156	2	0.078	8.959	<0.001	0.058
Experiment × condition	0.021	4	0.005	0.597	0.665	0.008
Error	2.528	290	0.009			

df: Degrees of Freedom.
F: F-statistic.
η_p^2 : Partial eta-squared.

TABLE 2 Repeated measures ANOVAs for the average hit rate as a function of Condition (low, medium, and high reward probability) for each experiment separately.

	Sum of squares	*df*	Mean square	*F*/*t*	*P*-value	η_p^2/d
Experiment 1						
Repeated measures ANOVA						
(Intercept)	102.09	1	102.09	2465.4	<0.001	
Error	2.029	49	0.041			
Condition	0.065	2	0.033	3.926	0.023	0.080
Error	0.814	98	0.008			
Mean comparisons						
Low versus medium				−1.866	0.068	0.264
Low versus high				−2.390	0.021	0.338
Medium versus high				−1.181	0.243	0.167
Experiment 2						
Repeated measures ANOVA						
(Intercept)	111.58	1	111.58	2395.5	<0.001	
Error	2.655	57	0.047			
Condition	0.072	2	0.036	3.476	0.034	0.057
Error	1.185	114	0.010			
Mean comparisons						
Low versus medium				−2.480	0.016	0.326
Low versus high				−1.783	0.08	0.234
Medium versus high				0.538	0.593	0.071
Experiment 3						
Repeated measures ANOVA						
(Intercept)	69.76	1	69.76	1056.5	<0.001	
Error	2.575	39	0.066			
Condition	0.046	2	0.023	3.360	0.040	0.079
Error	0.529	78	0.007			
Mean comparisons						
Low versus medium				−2.881	0.006	0.456
Low versus high				−1.393	0.172	0.220
Medium versus High				1.091	0.282	0.172

df: Degrees of Freedom.
F/t: F-statistic/t-statistic.
η_p^2 /d: Partial eta-squared/Cohen's d.

suggest interactions between two independent tasks performed in temporal proximity. Next, we turned to behavioral modeling to elucidate specific computational mechanism that may be touched by the ARR.

Modeling results

Parameter fitting and model selection was performed concurrently using a HBI method (Piray et al., 2019). To select a model, we utilized the PXP, which estimates the likelihood of a model providing the best explanation of the observed behaviors, as compared to all other considered models, while also taking into account the possibility that none of the compared models is supported by the data.

The PXPs for the tested models are shown in **Figure 2A**, with the largest PXP obtained for the "FB-R" model. In brief, the "FB-R" model presumes (i) that the ARR modulates the prediction error signal, and (ii) that manipulations of the ARR affects behavior *via* an accumulation of reward. Notably, the "FB-R" model obtained the highest PXP also when the analysis was repeated for each experiment separately (**Figure 2B**). For visualization purposes, the three parameters of the "FB-R" model are shown in **Figure 2C**. To demonstrate that these model parameters are meaningful (Wilson and Collins, 2019), we successfully recovered parameter values used to generate simulated behaviors (**Figure 2D**; see section "Materials and methods" for a description of this procedure).

One critical question is whether the "FB-R" model is capable of reproducing observed behaviors of interest. Before testing

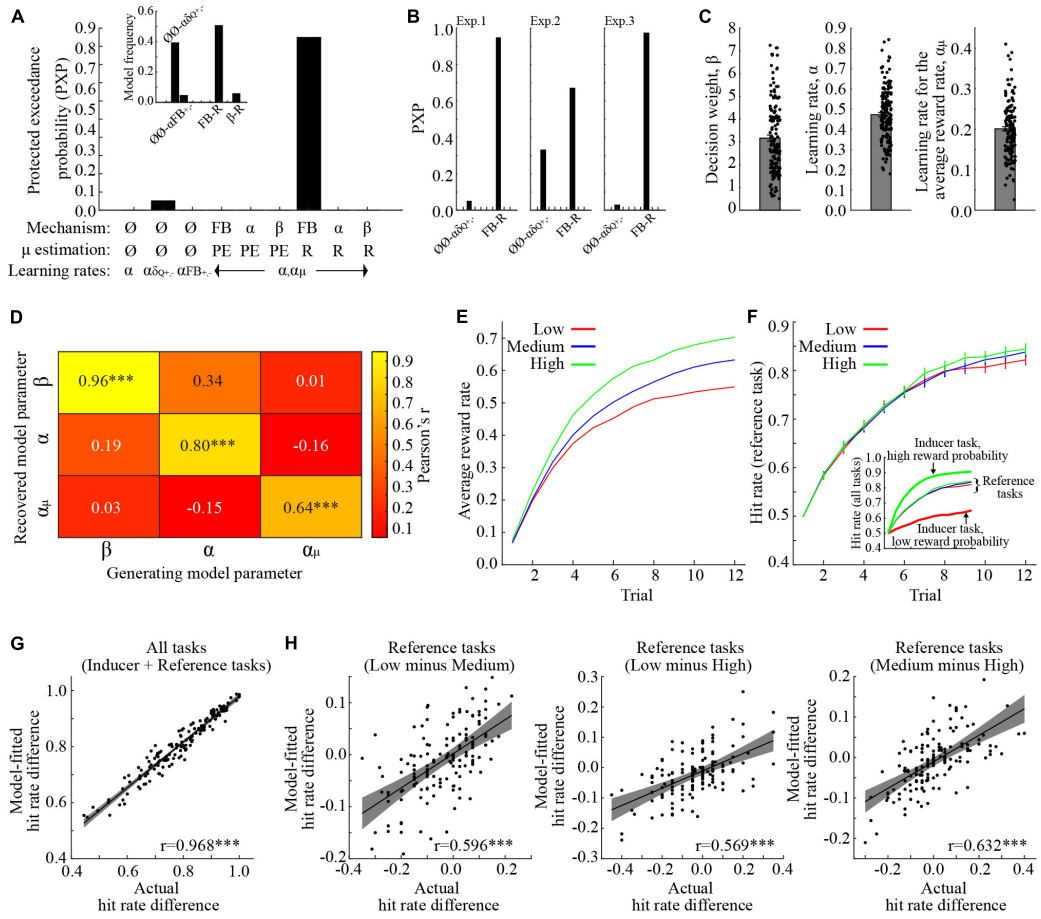

FIGURE 2
Behavioral modeling. **(A)** Protected exceedance probabilities (PXP) when model parameters were fitted across all experiments. The highest PXP was obtained for the FB-R model, where the average reward feedback influences the prediction error component of the model. The inset shows the model frequencies. **(B)** PXP's when model parameters were fitted in each experiment separately. In all experiments, the FB-R model had the highest PXP. **(C)** The model parameters, when fitted across experiments. **(D)** Correlation coefficients between model parameters used to generate behavior, and their recovered counterparts. **(E)** The average reward rate for the reference tasks in the different conditions. **(F)** Model-fitted hit rates for the reference tasks, collapsed across the three experiments. The inset also shows performance for the inducer tasks with low- and high reward probabilities. Inducer/reference tasks are, respectively denoted by bold / narrow lines. **(G)** Correlation between actual- and model-fitted learning performance, averaged across all conditions and tasks. **(H)** Correlations between actual- and model-fitted differences in learning performance for the reference task in the different conditions. $^{***}p < 0.001$. All error bars indicate the standard error of the mean.

this, we first display the "FB-R" model's estimate of average accumulated reward in each trial for the reference task in the different conditions (Figure 2E). As would be expected, the ARR increases when the inducer task provides higher reward probabilities. Model-fitted hit rates are shown in Figure 2F. To ensure that the selected model is capable of reproducing the same behavioral effects of interest, we entered model-fitted hit rates (averaged across trials 9–12) into the same mixed-effects ANOVA as for actual behavior. As with actual behavior, there was a significant main effect of Condition [F (2,290) = 7.217, $p < 0.001$, η_p^2 0.047, ANOVA], and even though the model suggests a main effect of Experiment [F (2,145) = 3.847, $p = 0.024$, η_p^2 0.050, ANOVA], there was no interaction between Experiment and Condition [F (4,290) = 0.170, p = 0.954, η_p^2 0.002, ANOVA]. For a full ANOVA table and comparisons between model-fitted means, see Table 3.

These results indicate that the model captures group-level behaviors. However, a good model should also be able to capture inter-individual differences in behavior. For this reason, we first averaged learning performance across the last third of the trials across all conditions, and found a positive and significant correlation between actual- and model fitted learning performance (Figure 2G; Pearson's $r = 0.968$, $p < 0.001$). Next, we correlated actual and model-fitted learning performances for our behavioral effects of interest, namely the differential hit rates for the reference task in the different conditions. All correlations were significant and positive [Low vs. Medium: Figure 2H, Pearson's $r = 0.596$, $p < 0.001$; Low vs. High: Figure 2I, Pearson's

$r = 0.569$, $p < 0.001$; Medium vs. High: **Figure 2J**, Pearson's $r = 0.632$, $p < 0.001$]. In other words, the "FB-R" model provides good fits to behavior, both on the group- and on the individual level.

In summary, we show that including the ARR in the model improves model fits. Moreover, the model with the most parsimonious fit presumes that ARR affects the prediction error signal *via* the accumulation of reward. These results extend the behavioral results by highlighting specific computational mechanism that is affected by manipulations of the ARR.

Discussion

In three separate tasks, we observed reduced learning performance in a "reference" task when it was interleaved with an "inducer" task that provided a low probability of reward. These results support the notion that affective information obtained in one task lingers and affects performance in other, temporally proximal tasks, even when these tasks are independent. These results extend previous research showing that the ARR affects task performance across trials within the same task (Guitart-Masip et al., 2011; Rigoli et al., 2016; Aberg et al., 2017, 2020). As such, these studies may suggest a general role for the ARR in enabling interactions between temporally proximal events.

What cognitive mechanism(s) may be responsible for such interactions? It has been suggested that the ARR affects intrinsic motivation, e.g., low ARR reduces intrinsic motivation (Niv, 2007; Niv et al., 2007; Hamid et al., 2016). The evolutionary advantage of such a mechanism is that, for example, it enables the ability to preserve energy when resources (e.g., food/water) are scarce *via* a reduced motivation to exert effort. Additionally, it may provide a signal which regulates foraging behaviors (Constantino and Daw, 2015). The present study suggests that motivational transfer occurs not only across trials within the same task, but also across different tasks. An interesting topic for future studies is to what extent a low ARR contributes to exploratory decisions or task-switching.

The present study defines ARR as rewards per trial, but another option is to calculate it as rewards per time (i.e., the opportunity cost of time; Niv et al., 2007). While the two definitions are highly correlated in the present study (because trial durations are more or less constant), this distinction may be particularly important in the context of associative memory formation because large inter-trial intervals, which causes a reduction in the ARR per time, was found to improve associative memory performance (Lattal, 1999; Gallistel and Gibbon, 2000). While this result seemingly contrasts with our previous results (Aberg et al., 2017, 2020), we observe that large inter-trial intervals affords additional memory processes that might act to enhance memory performance, e.g., memory rehearsal (Reitich-Stolero and Paz, 2019). Yet, similarities and differences between the behavioral impact of average reward *per time* and *per trial* need to be addressed in future studies. For example, to test for motivational transfer of ARRs *per time*, the present task could be modified with an inducer task that applies different inter-trial intervals to manipulate the ARR per time.

Goal-directed motivation may depend on three factors, namely outcome controllability, outcome value, and effort cost (Grahek et al., 2019). In the present study, the cost of exerting effort was arguably similar across the different conditions. In addition, because all behavioral effects were similar across experiments with different outcome values, an effect which could be explained by the contextual scaling of available rewards (Palminteri et al., 2015), the impact of outcome values on motivation was negligible. By contrast, the different reward probabilities associated with the inducer task in the different conditions may have influenced the perceived outcome controllability (e.g., perceived controllability was large / small when reward probabilities were large / small). In other words, while the present study set out to test the following chain of events: "Average reward rate in inducer task -> Motivation -> Learning in reference task" path, we may actually have tested another chain of events, namely "Average reward rate in inducer task -> Perceived control in inducer task -> Motivation -> Learning in reference task." Put simply, the reduced learning performance for the reference task in the condition with low reward probabilities may have been due to a transfer of reduced motivation induced by low perceived control in the inducer task. To test the motivational transfer of motivation induced by perceived control, an experiment could be conducted where the inducer task manipulates perceived control without altering the ARR. Such an experiment would test the chain "Perceived control inducer task -> Motivation -> Learning in independent task."

Low perceived controllability has been associated with a variety of anxiety-related disorders, such as generalized anxiety disorder, post-traumatic stress disorder, panic disorder, social anxiety disorder, and obsessive-compulsive disorder (Gallagher et al., 2014), as well as depression (White, 1959; Abramson et al., 1989), and an increased vulnerability to develop a mental disorder (Barlow, 2000; White et al., 2006). Accordingly, the present study may inform psychopathology by showing that perceived control reduces intrinsic motivation, and how this may transfer to affect other independent behaviors. A potential flip-side of this research is that increasing the perception of control may also transfer and boost other motivated behaviors. Yet, future studies are clearly needed to investigate potential interactions between ARRs and perceived control, as well as the link between such interactions and inter-individual differences in personality traits and mental disorders.

The behavioral modeling revealed that the impact of ARRs on behavior is due to an alteration of the prediction error signal, rather than affecting learning rates or the tendency to make more exploratory decisions. This result resonates with

TABLE 3 Repeated measures ANOVAs for the model-fitted average hit rate as a function of Condition (low, medium, and high reward probability).

	Sum of squares	*df*	Mean square	*F*/*t*	*P*-value	η_p^2/d
Mixed-effects ANOVA						
(Intercept)	292.13	1	292.13	7453.3	<0.001	
Experiment	0.302	2	0.151	3.847	0.024	0.050
Error	5.683	145	0.039			
Condition	0.036	2	0.018	7.217	<0.001	0.047
Experiment x Condition	0.002	4	<0.001	0.170	0.954	0.002
Error	0.720	290	0.003			
Mean comparisons						
Low versus medium				−2.185	0.031	0.180
Low versus high				−3.608	<0.001	0.297
Medium versus high				−1.771	0.079	0.146

df: Degrees of Freedom.
F/t: F-statistic/t-statistic.
η_p^2 /d: Partial eta-squared/Cohen's d.

our previous research showing that the ARR affected the neural correlates of feedback processing (Aberg et al., 2020), and suggestions that ARRs are coded in low tonic dopamine levels (Niv, 2007; Niv et al., 2007). However, recent evidence suggests that ARRs are associated with dopamine release in the nucleus accumbens (Hamid et al., 2016; Mohebi et al., 2019), but not with the activity of midbrain dopamine neurons (Mohebi et al., 2019). This seemingly surprising result, given that the nucleus accumbens receives dense projections from midbrain dopamine neurons (Ikemoto, 2007), indicates a complex relationship between ARRs, motivation, dopamine, and learning, which is still a hot topic for on-going research (for a recent insightful review, see Berke, 2018). Yet, because prediction errors are coded by midbrain dopamine neurons (Schultz, 2016) and the nucleus accumbens (Rutledge et al., 2010; Hart et al., 2014), it could be predicted that the impact of ARRs on reinforcement learning in the present tasks involves altered prediction error coding within these brain regions.

Behavioral models which included a factor that allowed the average rate to affect behavior consistently outperformed models which allowed no interaction between tasks. This result supports the notion of motivational transfers between tasks and trials, and therefore highlights the importance of including this interaction in future studies. For example, reinforcement learning tasks are sometimes presented in a block-wise fashion, but at other times in an inter-leaved fashion. Different models may therefore be needed to model behavior in these conditions in order to capture behavioral variability that can be attributed to motivational transfer between tasks. Another thing to consider is that while the ARR was here determined by the different feedbacks, it seems reasonable that an individual's perceived performance may also contribute to the ARR. For example, intuition dictates that performing a difficult or effortful task would eventually lead to a drop in motivation, as compared to easier tasks, even without the presentation of explicit performance feedback. Interestingly, in visual perceptual learning, interleaving trials of a difficult task with trials of an easier tasks impedes learning, even in the presence of performance feedbacks (Aberg and Herzog, 2009, 2010). Future studies need to address the relationship between externally and internally generated performance feedbacks and motivational transfer *via* the ARR.

Limitations

A first limitation is that both the inducer- and the reference task were reinforcement learning tasks, and therefore depended on the same neurocomputational mechanisms, e.g., prediction errors and the brain regions that code them. It therefore remains unknown whether our findings can be replicated using two different, e.g., two non-learning tasks or the combination of a learning- and a non-learning task. Even further, to what extent can ARRs obtained in a computerized game transfer and affect real-life motivated behaviors?

Second, we did not observe a difference in behavior between the conditions where the induced reward probability was medium and high, suggesting a non-linear relationship between ARRs and motivation. One explanation could be that the perceived difficulty/perceived controllability was similar for the medium and high average reward conditions (but different as compared to the low reward probability condition). Another option is that even if the difficulty/controllability is perceived as different, the motivational impact of medium and high reward probabilities could be similar. Supporting these notions, it was recently reported that the optimal rate of reward for binary classification learning was around 85% (Wilson et al., 2019), a value which is positioned rather in-between the medium (0.75) and the high (∼0.92) reward probabilities used in the present task. Yet, because we did not acquire any self-reported

assessments about task-difficulties, this issue remains a topic for future studies.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board (IRB) of the Weizmann Institute. Written informed consent was not provided because the experiment was conducted online, yet participation required providing informed consent by reading and agreeing to an online version of the consent form.

Author contributions

KA and RP designed the experiment and wrote the manuscript. KA collected and analyzed the data. Both authors contributed to the article and approved the submitted version.

Funding

KA was the incumbent of the Sam and Frances Belzberg Research Fellow Chair in Memory and Learning. The work was supported by a Joy-Ventures grant, an ISF #2352/19 and an ERC-2016-CoG #724910 grant to RP.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aberg, K. C., Doell, K. C., and Schwartz, S. (2015). Hemispheric asymmetries in striatal reward responses relate to approach-avoidance learning and encoding of positive-negative prediction errors in dopaminergic midbrain regions. *J. Neurosci.* 35, 14491–14500. doi: 10.1523/JNEUROSCI.1859-15.2015

Aberg, K. C., and Herzog, M. H. (2009). Interleaving bisection stimuli - randomly or in sequence - does not disrupt perceptual learning, it just makes it more difficult. *Vision Res.* 49, 2591–2598. doi: 10.1016/j.visres.2009.07.006

Aberg, K. C., and Herzog, M. H. (2010). Does perceptual learning suffer from retrograde interference? *PLoS One* 5:e14161. doi: 10.1371/journal.pone.0014161

Aberg, K. C., Kramer, E. E., and Schwartz, S. (2020). Interplay between midbrain and dorsal anterior cingulate regions arbitrates lingering reward effects on memory encoding. *Nat. Commun.* 11:1829. doi: 10.1038/s41467-020-15542-z

Aberg, K. C., Muller, J., and Schwartz, S. (2017). Trial-by-Trial modulation of associative memory formation by reward prediction error and reward anticipation as revealed by a biologically plausible computational model. *Front. Hum. Neurosci.* 11:56. doi: 10.3389/fnhum.2017.00056

Abramson, L. Y., Alloy, L. B., and Metalsky, G. I. (1989). Hopelessness depression - a theory-based subtype of depression. *Psychol. Rev.* 96, 358–372.

Barlow, D. H. (2000). Unraveling the mysteries of anxiety and its disorders from the perspective of emotion theory. *Am. Psychol.* 55, 1247–1263. doi: 10.1037//0003-066x.55.11.1247

Berke, J. D. (2018). What does dopamine mean? *Nat. Neurosci.* 21, 787–793. doi: 10.1038/s41593-018-0152-y

Constantino, S. M., and Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cogn. Affect. Behav. Neurosci.* 15, 837–853. doi: 10.3758/s13415-015-0350-y

Eldar, E., and Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nat. Commun.* 6:6149. doi: 10.1038/ncomms7149

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

Gallagher, M. W., Bentley, K. H., and Barlow, D. H. (2014). Perceived control and vulnerability to anxiety disorders: a meta-analytic review. *Cogn. Therapy Res.* 38, 571–584. doi: 10.1007/s10608-014-9624-x

Gallistel, C. R., and Gibbon, J. (2000). Time, rate, and conditioning. *Psychol. Rev.* 107, 289–344. doi: 10.1037/0033-295x.107.2.289

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition* 173, 34–42. doi: 10.1016/j.cognition.2017.12.014

Grahek, I., Shenhav, A., Musslick, S., Krebs, R. M., and Koster, E. H. W. (2019). Motivation and cognitive control in depression. *Neurosci. Biobehav. Rev.* 102, 371–381. doi: 10.1016/j.neubiorev.2019.04.011

Guitart-Masip, M., Beierholm, U. R., Dolan, R., Duzel, E., and Dayan, P. (2011). Vigor in the face of fluctuating rates of reward: an experimental examination. *J. Cogn. Neurosci.* 23, 3933–3938. doi: 10.1162/jocn_a_00090

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. *Nat. Neurosci.* 19, 117–126. doi: 10.1038/nn.4173

Hart, A. S., Rutledge, R. B., Glimcher, P. W., and Phillips, P. E. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *J. Neurosci.* 34, 698–704. doi: 10.1523/JNEUROSCI.2489-13.2014

Ikemoto, S. (2007). Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. *Brain Res. Rev.* 56, 27–78. doi: 10.1016/j.brainresrev.2007.05.004

Lattal, K. M. (1999). Trial and intertrial durations in pavlovian conditioning: issues of learning and performance. *J. Exp. Psychology-Animal Behav. Processes* 25, 433–450. doi: 10.1037/0097-7403.25.4.433

Lin, B., Cecchi, G. A., Bouneffouf, D., Reinen, J. M., and Rish, I. (2020). "A story of two streams: reinforcement learning models from human behavior and neuropsychiatry," in *Proceedings of the AAMAS*, Auckland.

Maia, T. V., and Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.* 14, 154–162. doi: 10.1038/nn.2723

Mohebi, A., Pettibone, J. R., Hamid, A. A., Wong, J. T., Vinson, L. T., Patriarchi, T., et al. (2019). Dissociable dopamine dynamics for learning and motivation. *Nature* 570, 65–70. doi: 10.1038/s41586-019-1235-y

Niv, Y. (2007). Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation? *Ann. N Y Acad. Sci.* 1104, 357–376. doi: 10.1196/annals.1390.018

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

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

Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., and Daw, N. D. (2019). Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLoS Comput Biol.* 15:e1007043. doi: 10.1371/journal.pcbi.1007043

Reitich-Stolero, T., and Paz, R. (2019). Affective memory rehearsal with temporal sequences in amygdala neurons. *Nat. Neurosci.* 22, 2050–2059. doi: 10.1038/s41593-019-0542-9

Rigoli, F., Chew, B., Dayan, P., and Dolan, R. J. (2016). The dopaminergic midbrain mediates an effect of average reward on pavlovian vigor. *J. Cogn. Neurosci.* 28, 1303–1317. doi: 10.1162/jocn_a_00972

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *Neuroimage* 84, 971–985. doi: 10.1016/j.neuroimage.2013.08.065

Rutledge, R. B., Dean, M., Caplin, A., and Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *J. Neurosci.* 30, 13525–13536. doi: 10.1523/Jneurosci.1747-10.2010

Schultz, W. (2016). Dopamine reward prediction error coding. *Dial. Clin. Neurosci.* 18, 23–32.

Sutton, R., and Barto, A. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press.

Watkins, C. J. C. H., and Dayan, P. (1992). Q-Learning. *Mach. Learn.* 8, 279–292.

White, K. S., Brown, T. A., Somers, T. J., and Barlow, D. H. (2006). Avoidance behavior in panic disorder: the moderating influence of perceived control. *Behav. Res. Ther.* 44, 147–157. doi: 10.1016/j.brat.2005.07.009

White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychol. Rev.* 66, 297–333. doi: 10.1037/h0040934

Wilson, R. C., and Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife* 8:e49547. doi: 10.7554/eLife.49547

Wilson, R. C., Shenhav, A., Straccia, M., and Cohen, J. D. (2019). The eighty five percent rule for optimal learning. *Nat. Commun.* 10:4646. doi: 10.1038/s41467-019-12552-4