

Reward-Mediated, Model-Free Reinforcement-Learning Mechanisms in Pavlovian and Instrumental Tasks Are Related

Neema Moin Afshar,¹  François Cinotti,² David Martin,³  Mehdi Khamassi,⁴ Donna J. Calu,^{3,5} Jane R. Taylor,^{1,6} and  Stephanie M. Groman^{1,7,8}

¹Department of Psychiatry, Yale School of Medicine, New Haven, Connecticut 06511, ²Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, United Kingdom, ³Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland 21201, ⁴Institute of Intelligent Systems and Robotics, Centre National de la Recherche Scientifique, Sorbonne University, 75005 Paris, France, ⁵Program in Neuroscience, University of Maryland School of Medicine, Baltimore, Maryland 21201, ⁶Department of Psychology, Yale University, New Haven, Connecticut 06520, ⁷Department of Neuroscience, University of Minnesota Medical School, Minneapolis, Minnesota 55455, and ⁸Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455

Model-free and model-based computations are argued to distinctly update action values that guide decision-making processes. It is not known, however, if these model-free and model-based reinforcement learning mechanisms recruited in operationally based instrumental tasks parallel those engaged by pavlovian-based behavioral procedures. Recently, computational work has suggested that individual differences in the attribution of incentive salience to reward predictive cues, that is, sign- and goal-tracking behaviors, are also governed by variations in model-free and model-based value representations that guide behavior. Moreover, it is not appreciated if these systems that are characterized computationally using model-free and model-based algorithms are conserved across tasks for individual animals. In the current study, we used a within-subject design to assess sign-tracking and goal-tracking behaviors using a pavlovian conditioned approach task and then characterized behavior using an instrumental multistage decision-making (MSDM) task in male rats. We hypothesized that both pavlovian and instrumental learning processes may be driven by common reinforcement-learning mechanisms. Our data confirm that sign-tracking behavior was associated with greater reward-mediated, model-free reinforcement learning and that it was also linked to model-free reinforcement learning in the MSDM task. Computational analyses revealed that pavlovian model-free updating was correlated with model-free reinforcement learning in the MSDM task. These data provide key insights into the computational mechanisms mediating associative learning that could have important implications for normal and abnormal states.

Key words: computational psychiatry; decision-making; incentive salience; model-based learning; model-free learning

Significance Statement

Model-free and model-based computations that guide instrumental decision-making processes may also be recruited in pavlovian-based behavioral procedures. Here, we used a within-subject design to test the hypothesis that both pavlovian and instrumental learning processes were driven by common reinforcement-learning mechanisms. Sign-tracking and goal-tracking behaviors were assessed in rats using a pavlovian conditioned approach task, and then instrumental behavior was characterized using an MSDM task. We report that sign-tracking behavior was associated with greater model-free, but not model-based, learning in the MSDM task. These data suggest that pavlovian and instrumental behaviors may be driven by conserved reinforcement-learning mechanisms.

Received June 9, 2022; revised Oct. 3, 2022; accepted Oct. 6, 2022.

Author contributions: D.J.C. and S.M.G. designed research; N.M.-A. and S.M.G. performed research; J.R.T. contributed unpublished reagents/analytic tools; N.M.-A., F.C., D.M., M.K., D.J.C., and S.M.G. analyzed data; and N.M.-A. and S.M.G. wrote the paper.

This work was supported by National Institutes of Health–National Institute on Drug Abuse Grants DA041480 (J.R.T.), DA043443 (J.R.T.), DA051598 (S.M.G.), and DA043533 (D.J.C.); McKnight Foundation Memory and Cognitive Disorders Award (D.J.C.); and the State of Connecticut Department of Mental Health and Addiction Services through its support of the Ribicoff Laboratories. We thank Matthew Roesch for leading discussions that made this collaborative work a possibility.

The authors declare no competing financial interests.

Correspondence should be addressed to Stephanie M. Groman at sgroman@umn.edu.

<https://doi.org/10.1523/JNEUROSCI.1113-22.2022>

Copyright © 2023 the authors

Introduction

Cues in the environment that predict rewards can acquire incentive value through pavlovian mechanisms (Flagel et al., 2009) and are necessary for the survival of an organism by facilitating predictions about biologically relevant events that enable an organism to engage in appropriate preparatory behaviors. Pavlovian incentive learning, however, can imbue cues with strong incentive motivational properties that exert control over behavior, which can lead to maladaptive and detrimental behaviors (Saunders and Robinson, 2013). For example, cues that are associated with drug use can enhance craving in addicts and because of

their control over behavior may precipitate relapse to drug-taking behaviors in abstinent individuals (Hammersley, 1992). Understanding the biobehavioral mechanisms underlying associative learning could, therefore, provide critical insights into how stimuli gain incentive salience.

Pavlovian associations have largely been presumed to occur through model-free, or stimulus–outcome, learning; cues that are predictive of rewards incrementally accrue value through a temporal-difference signal that is likely to be mediated by mesolimbic dopamine (Huys et al., 2014; Nasser et al., 2017; Saunders et al., 2018). Theoretical work has, however, proposed that pavlovian associations may also involve learning that is described in the computational field as model based (MB; Dayan and Berridge, 2014; Lesaint et al., 2014), whereby individuals learn internal models of the statistics of action–outcome contingencies. This hypothesis has been supported by data indicating that pavlovian associations not only represent accrued value but also the identity of pavlovian outcomes (Robinson and Berridge, 2013) and by neuroimaging studies that identify neural signatures of model-free and model-based learning in humans during a pavlovian association task (Wang et al., 2020).

Pavlovian autoshaping procedures have been used to quantify the extent to which animals attribute incentive salience to cues predictive of rewards (Flagel et al., 2009, 2011; Nasser et al., 2015). When animals are presented with a cue associated with food reward delivery, the majority of rats known as sign trackers (STs) will approach and interact with the cue, whereas other rats known as goal trackers (GTs) will approach the location of the reward delivery (Hearst and Jenkins, 1974; Boakes, 1977). Rats that display sign-tracking behaviors, therefore, attribute incentive salience to the cue, whereas rats that display goal-tracking behaviors do not (Robinson and Flagel, 2009), or at least acquire less incentive to the cue than the goal. Our computational work (Lesaint et al., 2014; Cinotti et al., 2019) has suggested that these conditioned responses may be linked to individual differences in the extent to which rats use model-free and model-based reinforcement-learning systems to guide their behavior. For example, when using a hybrid reinforcement-learning model to simulate pavlovian approach behaviors we were able to recapitulate sign-tracking behaviors by increasing the weight of model-free updating and, notably, goal-tracking behaviors by increasing the weight of model-based updating (Cinotti et al., 2019). Variation in pavlovian approach behaviors in rodents may, therefore, reflect individual differences in model-free and model-based control over behavior (Dayan and Berridge, 2014; Lesaint et al., 2014).

Use of the multistage decision-making (MSDM) task in humans (Daw et al., 2011; Culbreth et al., 2016) and animals (Miller et al., 2017; Groman et al., 2019a; Akam et al., 2021) has provided empirical evidence that instrumental behavior is influenced by both model-free and model-based reinforcement learning computations. It is not known, however, if the relative contribution of model-free and model-based mechanisms that are recruited in an individual during pavlovian autoshaping procedures are predictive of their relative contribution during instrumental procedures, such as in the MSDM task (Sebold et al., 2016). If true, this could suggest that the computational mechanisms underlying learning are not unique to pavlovian or instrumental mechanisms but may represent a common reinforcement-learning framework within the brain that could be useful for restoring the learning mechanisms that are abnormal in disease states (Doñamayor et al., 2021; Groman et al., 2021).

In the current study we sought to test the hypothesis that ST rats would preferentially employ a model-free strategy in an instrumental task, whereas GT rats would preferentially employ a model-based strategy. Pavlovian conditioned approach was assessed in rats (Keefer et al., 2020) before model-free and model-based reinforcement-learning was assessed in a rodent analog of the MSDM task (Groman et al., 2019a). We report that sign-tracking behavior is correlated with individual differences in reward-mediated model-free, but not model-based, learning in the MSDM task. These data suggest that the model-free reinforcement-learning systems recruited during pavlovian conditioning parallel those recruited in instrumental learning.

Materials and Methods

Subjects

Twenty male Long-Evans rats were purchased from Charles River Laboratories at ~6 weeks of age. Rats were pair housed in a climate-controlled vivarium on a 12 h light/dark cycle (lights on at 7:00 A.M., lights off at 7:00 P.M.). Rats had *ad libitum* access to water and underwent dietary restriction to 90% of their free-feeding body weight throughout the experiment to maintain the same hunger state in both the pavlovian and instrumental environments. Experimental procedures were approved by the Institutional Animal Care and Use Committee at Yale University and were in accordance with the National Institutes of Health institutional guidelines and *Public Health Service Policy on Humane Care and Use of Laboratory Animals*.

Pavlovian conditioned approach

Rats were first trained using a pavlovian conditioned approach task as previously described (Keefer et al., 2020). During a single trial, a retractable lever, conditioned stimulus (CS), located to the left or right of a food cup was inserted into the chamber for 10 s. As the lever retracted, a single sucrose pellet (45 mg; BioServ) was dispensed into the food cup. This CS and unconditioned stimulus (US) pairing occurred on a variable-interval 60 s schedule, and each CS-US pairing was present 25 times in each session. Each rat underwent a single daily session on the pavlovian conditioned approach task for 5 consecutive days. The primary dependent measures collected were latency to approach the lever and food cup as well as the number and probability of interactions rats made with the lever and food cup within each session. These dependent measures were used to generate a pavlovian score for each session a rat completed (see below, Data analysis). This pavlovian score is typically referred to as the Pavlovian Conditioned Approach (PCA) score; however, to avoid confusion with the data reduction technique known as principal component analysis (also commonly referred to as a PCA) we refer to this measure as the PavCA score.

Deterministic MSDM task

Following the pavlovian conditioning approach sessions, rats were trained to make operant responses (e.g., nose pokes and lever responses) to receive a liquid reward delivery (90 μ l of 10% sweetened condensed milk) in a different environment than that used for the pavlovian conditioning approach task. Once operant responding had been established, rats were trained on a deterministic MSDM task using procedures previously described (Groman et al., 2019a). In the deterministic MSDM task, choices in the first stage deterministically led to the second-stage state. Second-stage choices were probabilistically rewarded. Rats initiated trials by making a response into the illuminated food cup. Two levers located on either side of the food cup were extended into the box and cue lights above the levers illuminated (s_a). A response made on one lever (s_{a1}) resulted in the illumination of two nose port apertures (e.g., ports 1 and 2, s_B), whereas a response made on the other lever (s_{a2}) would result in the illumination of two other apertures (ports 3 and 4, s_C). Entries into either of the illuminated apertures were probabilistically reinforced using an alternating block schedule.

Each rat was assigned to one specific lever-port configuration (configuration 1, left lever→port 1, 2; right lever→port 3, 4; configuration 2, left lever→port 3, 4; right lever→port 1, 2), which was maintained

through the study. Reinforcement probabilities assigned to each port, however, were pseudorandomly designated at the beginning of each session (0.90 vs 0.10 or 0.40 vs 0.15; see Fig. 4A), terminated when 300 trials had been completed or 90 min had elapsed, whichever occurred first. Trial-by-trial data were collected for individual rats, and the probability that rats would select the first-stage option leading to the highest reinforced second-stage option [$p(\text{correct}|\text{stage1})$] was calculated, as well as the probability that rats would select the highest reinforced second-stage option [$p(\text{correct}|\text{stage2})$].

Training on the deterministic MSDM task occurred for the following primary reasons: (1) reduce spatial biases, which are common in rats; (2) ensure rats understood the alternating probabilities of reinforcement at the second-stage options; and (3) verify that rats understood the structure of the task and how first-stage choices led to different second-stage options. If rats appreciated the reinforcement probabilities assigned to the second-stage options and how choices in the first stage influence the availability of second-stage options, then the probability that rats choose the first-stage option leading to the second-stage option with the maximum reward probability [e.g., $p(\text{correct} | \text{stage 1})$] should be significantly greater than that predicted by chance. Rats were trained on the deterministic MSDM until they met the criteria of a $p(\text{correct}|\text{stage1})$ being significantly greater than chance in four of the five sessions after completing the 35th training session on the deterministic MSDM. If rats did not meet the criterion after completing 43 sessions on the deterministic MSDM ($N = 3$), training was terminated regardless of $p(\text{correct}|\text{stage1})$.

Probabilistic MSDM task

Choice behavior was then assessed in the probabilistic MSDM task. Initiated trials resulted in the extension of two levers and illumination of cue lights located above each lever. For most trials (70%), first-stage choices led to the illumination of the same second-stage state that was deterministically assigned to that first-stage choice in the deterministic MSDM (referred to as a common transition). On a limited number of trials (30%), first-stage choices led to the illumination of the second-stage state most often associated with the other first-stage choice (referred to as a rare transition). Second-stage choices were probabilistically reinforced using the same alternating block schedule as that of the deterministic MSDM task. Rats completed 300 trials across five daily sessions on the probabilistic MSDM task.

Trial-by-trial data (~1500 trials/rat) were collected to conduct logistic regression analyses of decision-making (described below). One rat was excluded from all analyses because of an extreme bias in the first-stage choice (e.g., rat chose one lever on 97% of all trials, regardless of previous trial events).

Data analysis

Pavlovian conditioned approach. To quantify the degree to which individual rats display sign-tracking or goal-tracking behaviors, a PavCA score was calculated for individual rats by averaging three standardized measures, collected as previously described (Meyer et al., 2012). These three measures were (1) a latency score, which was the average latency to make a food cup response during the CS, minus the latency to lever press during the CS, divided by the duration of the CS (10 s); (2) a probability score, which was the probability that the rat would make a lever press, minus the probability that the rat would make a food cup response across the session; and (3) a preference score, which was the number of lever contacts during the CS, minus the number of food cup contacts during the CS, divided by the sum of these two measures. The PavCA score ranged between -1.0 and 1.0 , with values closer to 1.0 reflecting a greater prevalence of ST behaviors and values closer to -1.0 reflecting a greater prevalence of goal-tracking behaviors. Previous studies have calculated the average PavCA score from the last two pavlovian sessions rats complete to classify rats as either exhibiting high or low ST behaviors (Morrison et al., 2015; Rode et al., 2020) as goal-tracking behaviors are less commonly observed within the population. We refer to this average measure as the summary PavCA score. We conducted a similar median split of the distribution of summary PavCA scores and classified rats as either exhibiting high sign-tracking behaviors (high ST rats; $N = 10$) or low sign-tracking behaviors (low ST rats; $N = 10$). All group-level

analyses reported in the current study were conducted using this binary classification.

Additionally, a trial-by-trial pavlovian score was calculated to serve as the dependent measure used in the computational analyses described below. Latency, probability, and preference scores were calculated on each trial, and the average of these measures was used to categorize an individual trial as approach to the lever or approach to the magazine. Specifically, the latency score was the average latency to make a food cup response, minus the latency to make a lever press during the CS, divided by the duration of the CS on that trial. The probability score was the probability that rats would make a lever press (+1) versus a food cup response (−1) on that trial. The preference score was the number of lever contacts during the CS, minus the number of food cup contact during the CS, divided by the sum of these measures for that trial. Although rats could vacillate between these responses within a single trial (e.g., approach lever, check magazine, approach lever), characterizing this within-trial variability is difficult and beyond the scope of the current study. The PavCA score and the trial-by-trial pavlovian score for each of the five pavlovian sessions was positively correlated (all R^2 values > 0.94 ; all p values < 0.001) suggesting that these measures were capturing the same variability in pavlovian approach behaviors.

Model-free and model-based learning in the pavlovian conditioned approach task. We have previously reported that individual differences in pavlovian approach behavior can be recapitulated using a combination of model-free and model-based reinforcement-learning algorithms (Lesaint et al., 2014; Cinotti et al., 2019). We sought to use this hybrid reinforcement-learning model to index the contribution of these reinforcement-learning systems to the pavlovian conditioned approach behaviors measured in the current study. This model combines the outputs of these two reinforcement-learning systems to determine the likelihood that rats will approach the lever or approach the magazine on each trial. The structure of each trial of the task is represented by a Markovian Decision Process (MDP) consisting of six different states (Fig. 1A) defined by the experimental conditions, such as the presence of the lever or of the food and the current position of the rat (e.g., close to the magazine or the lever). The five different actions are explore the environment (goE), approach the lever (goL), approach the magazine (goM), engage the closest stimuli or $\text{eng}_{\langle L \rangle / \langle M \rangle}$, and eat the reward, and state transitions given a selected action are deterministic. For example, if a rat in state 1 (s_1) chooses the action goL, it will always lead to state 2 (s_2) whereas if a rat in state 1 (s_1) chooses the action goM, it will lead to state 3 (s_3). Action values for all possible actions in the current state are generated by the decision-making model, which consists of both an MB and a feature model-free (FMF) reinforcement-learning algorithm (Fig. 1B). The MB and FMF value estimates are combined as a sum into a weighted value determined by the parameter ω . An ω parameter closer to 1 indicates that action values are more influenced by the MB computations, whereas an ω parameter closer to zero indicates that action values are more influenced by the FMF computations. The weighted values are fed into a softmax function representing the action selection mechanism.

The FMF system, compared with instrumental reinforcement-learning algorithms, assigns value representations to the features associated with each action rather than to the states of the task, which allows a generalization of values among different states. For example, when the rat goes toward the magazine in state 1 (s_1) or engages the magazine in state 3 (s_3), it does so motivated by the same feature value [$V(M)$] in these two different states, which means $V(M)$ is updated twice in the course of a single trial. After each action, the value of the corresponding feature is updated according to a standard temporal difference (TD) learning rule by first computing a reward prediction error (δ) as follows:

$$\delta = r - V(f(s_t, a_t)),$$

where r is equal to 1 or 0 if reward delivery occurs or not, respectively. This reward prediction error is integrated in the current estimate of the value of the chosen feature as follows:

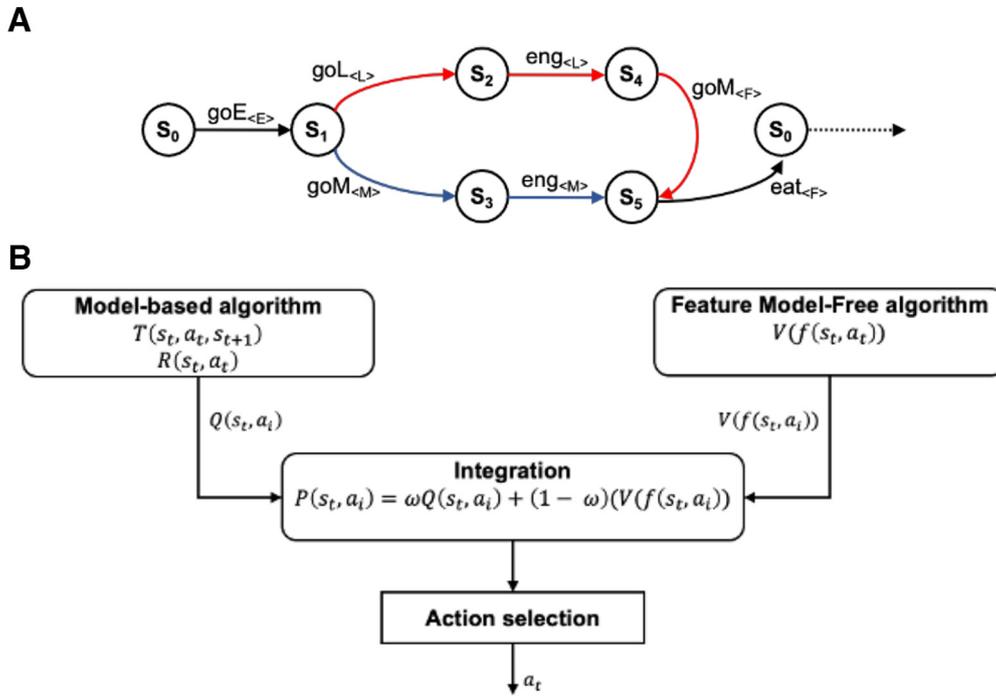


Figure 1. The FMF-MB decision-making model. **A**, The Markov decision process of a single trial from the pavlovian approach task. The following five possible actions lead deterministically from one state to the next: goE, goL, goM, eng, and eat the reward. Each of these actions focuses on a specific feature indicated in angle brackets, E, L, M, and F. These are the features used by the FMF learning component. The red path corresponds to sign-tracking behavior and the blue path to goal-tracking behavior. **B**, Schematic of the FMF-MB decision-making model adapted from Lesaint et al. (2014) and Ginotti et al. (2019). The model combines an MB learning system, which learns the structure of the MDP and then calculates the relative advantage of each action in a given state, and an FMF system, which attributes a value to different features of the environment that is generalized across states (e.g., the same value of the magazine is used in states 1 and 4). The advantage function and value function are weighted by ω , their relative importance determining the sign- versus goal-tracking tendency of the individual and then passed to the action selection mechanism modeled by a softmax function.

$$V(f(s_t, a_t)) \leftarrow V(f(s_t, a_t)) + \alpha \delta,$$

with the learning rate (e.g., α) bounded between 0 and 1. In contrast to our previous FMF algorithm, the discounting parameter γ was not included here. This was because our model comparisons (see below, Results) indicated that this parameter was not explaining unique variance in the approach behavior of this group of rats.

The TD learning rule was only applied to the selected feature (E, environment; L, lever; M, magazine) in each state transition, except in the case of food, which was equal to 1, the value of reward. Because the rat is likely to visit the magazine during the intertrial interval (ITI), the values of the magazine are revised between state 5 and state 0 according to the following equation:

$$V(M) \leftarrow (1 - u_{ITI}) \times V(M),$$

where u_{ITI} determines the rate at which action values for the magazine $[V(M)]$ decay during ITI.

The MB system relies on learned transition T and reward R functions for updating action values. The transition value function aims to determine the probability of going from one state to the next given a certain action. After transitioning from state s_t to s_{t+1} by performing action a_t , the transition $T(s_t, a_t, s_{t+1})$ is updated according to the following:

$$T(s_t, a_t, s_{t+1}) \leftarrow (1 - \alpha)T(s_t, a_t, s_{t+1}) + \alpha,$$

with initial values of T set to 0 for all possible state and action combinations. The T values for unvisited states are decreased according to the following:

$$T(s_t, a_t, s_{t+1}) \leftarrow (1 - \alpha)T(s_t, a_t, s_{t+1}).$$

Because the environment is deterministic, $T(s, a, s)$ should converge perfectly toward values of 1 for all possible state transitions and remain

at a value of 0 for all impossible state transitions (e.g., $s_1 \rightarrow s_4$). Similarly, the reward function $R(s, a)$ is updated according to the following:

$$R(s_t, a_t) \leftarrow (1 - \alpha)R(s_t, a_t) + ar,$$

where r is set to 1 for (s_5 , eat) and 0 otherwise. Initially, $R(s, a)$ is equal to 0 for all state-action pairs. Across actions and experience in each state, $R(s, a)$ will converge to a value of 1 for (s_5 , eat) state-action pair and remain at a value of 0 for all other state-action pairs. The action-value functions for each possible action a_i in the current state s_t are then calculated according to the following:

$$Q(s_t, a_i) = R(s_t, a_i) + \gamma \sum_j T(s_t, a_i, s_j) \max_k Q(s_j, a_k),$$

where γ is the discounting parameter. Once the FMF and MB systems have outputted the feature values and the advantages of the possible actions, these are integrated through a weighted sum as follows:

$$P(s_t, a_i) = \omega Q(s_t, a_i) + (1 - \omega)V(f(s, a_i)),$$

with ω bounded between 0 and 1. These integrated values are then entered into a softmax function to compute the probability of selecting each action as follows:

$$p(a_t = a_i) = \frac{e^{\beta P(s_t, a_i)}}{\sum_j e^{\beta P(s_t, a_j)}},$$

where β is the inverse temperature parameter quantifying choice stochasticity.

This model contained five free parameters, the learning rate α , the discounting factor γ , the inverse temperature β , the ITI update factor

u_{ITI} , and the integration factor ω . Trial-by-trial behavior was classified as either being a sign-tracking or goal-tracking behavior and fit with five free parameters selected to maximize the likelihood of sequence of behavior of each rat as follows:

$$L = \sum_{\text{trials}} \ln(P(a_t | \alpha, \beta, \gamma, \omega, u)).$$

To avoid local maxima, starting values for each free parameter were optimized using a grid search so that each parameter had three possible initial values, and all 3^5 possible combinations were tested as the starting point of the gradient descent procedure to maximize the likelihood L . Each pavlovian session only contained 25 trials, which proved to be difficult for obtaining reliable parameter estimates from this reinforcement-learning model. To improve the accuracy and reliability of parameter estimates and model fit, trial-by-trial data for all five of the pavlovian sessions were concatenated into a single dataset for individual rats and analyzed with the reinforcement-learning model. The parameter estimates that yielded the largest log likelihood were retained and are reported in Table 1.

Logistic regression of choice data in the MSDM tasks. We have previously shown that the choice behavior of rats in the MSDM task is guided by previous trial events, such as previous trial outcome, choice, and, in the probabilistic MSDM task, state transitions. Trial-by-trial choice data in the deterministic and probabilistic MSDM was analyzed with a logistic regression model using the `glmfit` function in MATLAB version 2017a (MathWorks). These logistic regression models predicted the likelihood that rats would select the same first-stage choice on the current trial (trial t) that they had on the previous trial (trial $t-1$), namely the probability of staying or $p(\text{stay})$. The model used to analyze choice data in the deterministic MSDM contained the following predictors.

- Intercept: +1 for all trials, which quantifies the tendency for rats to repeat the same first stage option regardless of any other trial events.
- Correct: +1 for trials where the rat selects the first stage option with a common transition leads to the highest reinforced stage 2 option.
–1 for trials where the rat selects the first stage option with a common transition leads to the lowest reinforced stage 2 option.
- Outcome: +1 if the previous trial resulted in a rewarded outcome.
–1 if the previous trial resulted in an unrewarded outcome.

The model used to analyze choice data in the probabilistic MSDM contained the same predictors as described above as well as two additional predictors.

- Transition: +1 if the previous trial included a common transition.
–1 if the previous trial included a rare transition.
- Transition-by-Outcome: +1 if the previous trial included a common transition and was rewarded or if it included a rare transition and was unrewarded.
–1 if the previous trial included a rare transition and was rewarded or included a common transition and was unrewarded.

The correct predictor in the logistic regression prevents spurious loading onto the transition-by-outcome interaction predictor (Akam et al., 2015), which can occur when using blocked schedules of reinforcement in the MSDM task. We included the correct predictor in all logistic regression models to ensure consistency across analyses and MSDM tasks. Critically, the regression coefficient applied to outcome quantifies model-free behavior, and the regression coefficient applied to the transition-by-outcome interaction quantifies model-based behavior.

Logistic regression of rewarded and unrewarded outcomes. We found that individual differences in the summary PavCA score was related to variation in the outcome regression coefficient (see below, Results). To

Table 1. Parameter estimates from the full hybrid model

Percentile	α	γ	β	u_{IT}	ω
25th	0.88	1	36.05	0.33	0.90
Median	0.26	0.90	5.11	0.12	0.72
75th	0.17	0.56	4.16	0.04	0.20

Values presented are those from the 25th, median, and 75th percentiles.

determine whether this relationship was because of differences in the influence of rewarded and/or unrewarded outcomes on choice behavior, we analyzed choice data in the MSDM task using a different logistic regression model that estimated the likelihood that rats would repeat the same first-stage choice based on whether the previous trial was rewarded or unrewarded. This logistic regression model, unlike the first, permitted an independent analysis of how each trial outcome (rewarded or unrewarded) influenced first-stage choices. The predictors included in this model were as follows.

- Intercept: +1 for all trials. This quantifies the tendency for rats to repeat the same first-stage option regardless of any other trial events.
- Rewarded: +1 if the previous trial was rewarded and the rat chose the same lever (first-stage choice) that was selected on the subsequent trial.
–1 if the previous trial was rewarded and the rat chose a different lever (first-stage choice) than what was selected on the subsequent trial.
0 if the previous trial was unrewarded.
- Unrewarded: +1 if the previous trial was unrewarded and the rat chose the same lever that was selected on the subsequent trial.
–1 if the previous trial was unrewarded and the rat chose a different lever than what was selected on the subsequent trial.
0 if the previous trial was rewarded.

Positive regression coefficients for the rewarded and unrewarded predictors indicate that rats are more likely to persist with the same first-stage choice, whereas negative regression coefficients indicate that rats are more likely to shift their first-stage choice. The probability that rats would repeat the same first-stage choice following rewarded and unrewarded trials was also calculated to examine how this more traditional measure of win-stay and lose-stay behaviors might differ between high and low ST rats.

Statistical analyses

Values presented are mean \pm SEM, unless otherwise noted. Statistical analyses were conducted in IBM SPSS (version 26), MATLAB (version 2017a, MathWorks), and R (<https://www.R-project.org>) software. Generalized linear models (GLMs; R `glmfit` package) were used to analyze the relationship between the summary pavlovian score and choice behavior of rats in the MSDM task. The dependent variable was a binary array coding for whether the first-stage choice was the same (+1) or different (0) from the previous trial. Predictors in the model could be correct, outcome, transition, transition-by-outcome interaction, and summary PavCA score or the binary classification of low ST or high ST rats. All higher-order (e.g., summary PavCA score times outcome times transition) and lower order (e.g., summary PavCA score times outcome) interactions were included in the model. Significant interactions were tested using progressively lower-order analyses. Another GLM was used to examine the relationship between the summary PavCA score and the influence of rewarded and unrewarded outcomes on first-stage choices. The dependent variable was a binary array coding for the first-stage choice (+1 for left lever and 0 for right lever). Predictors in the model were reward, unrewarded, and summary pavlovian score. All interactions (e.g., summary PavCA score times rewarded) were included in the model and significant interactions tested using lower-order analyses.

All other analyses were performed in SPSS. Repeated-measures data were entered into a generalized estimating equation model using a

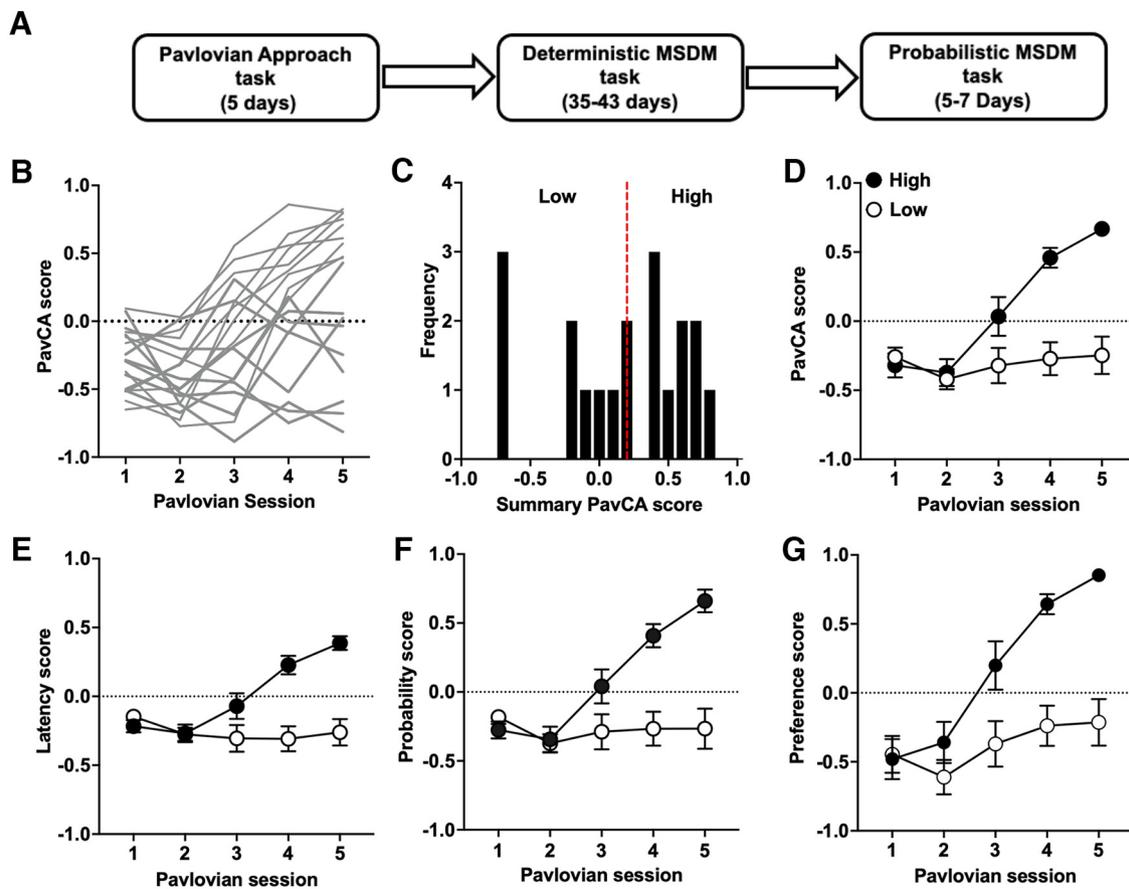


Figure 2. Pavlovian approach task. **A**, Schematic of the experimental design. Rats ($N = 20$) underwent five sessions on the pavlovian approach task before being trained in the deterministic MSDM task (35–43 d) and tested in the probabilistic MSDM tasks (5–7 d). **B**, The PavCA score for individual rats across the five pavlovian sessions. **C**, Distribution of the summary PavCA score obtained from pavlovian sessions 4 and 5. Rats were divided into two groups based on a median split (red line) of the summary PavCA score; low ST rats ($N = 9$) and high ST rats ($N = 10$). **D**, The PavCA score for low ST and high ST rats across the five pavlovian sessions. **E**, The latency score increased in high ST rats across the pavlovian sessions but did not change in the low ST rats. **F**, The probability score increased in the high ST rats across the pavlovian sessions but did not change in the low ST rats. **G**, Preference score increased in the high ST rats across the pavlovian sessions but did not change in the low ST rats. Values presented are mean \pm SEM.

probability distribution based on the known properties of these data. Specifically, event data (e.g., number of trials in which rats chose the highest reinforced first-stage option) were analyzed using a binary logistic distribution. Relationships between dependent variables (e.g., ω and model-free learning) were tested using Spearman's rank correlation coefficient.

Results

Pavlovian conditioned approach

Pavlovian incentive learning was assessed in rats in a pavlovian conditioned approach task for 5 d (Fig. 2A,B). The summary PavCA score was calculated and a median split conducted to classify rats as exhibiting either high ($N = 10$) or low ($N = 9$) sign-tracking behaviors (Fig. 2C). As expected, the PavCA score increased across the sessions in the high ST group (Wald $\chi^2 = 91.33$; $p < 0.001$) but not in the low ST group (Wald $\chi^2 = 0.23$; $p = 0.63$; Fig. 2D). We then examined how lever- and food-cup-directed behaviors changed across the five pavlovian conditioning sessions in both high and low ST rats (Fig. 2E–G). *Post hoc* analysis of the group (high vs low ST) \times session interaction (Wald $\chi^2 = 30.37$; $p < 0.001$) indicated that the latency score, the probability score, and the preference score increased across the pavlovian sessions in the high ST group (Wald $\chi^2 = 68.28$; $p < 0.001$) but not in the low ST group (all Wald χ^2 values < 0.99 ; $p > 0.32$). These session-dependent changes in the high ST

rats are similar to observations that we, and others, have reported using pavlovian conditioned approach tasks (Flagel et al., 2011; Saunders and Robinson, 2011; Keefer et al., 2020).

Computational analysis of pavlovian approach behavior

Each pavlovian session consisted of only 25 trials, which limited our ability to obtain reliable and accurate reinforcement-learning parameter estimates for each session and each rat. To overcome this, we concatenated the trial-by-trial data from all five pavlovian sessions into a single 125-trial dataset for individual rats and fitted these data with the hybrid model described above, and estimates of the five parameters (e.g., α , β , γ , u_{ITI} , ω) are presented in Table 1. We also compared the fits of this hybrid model to other variants of this model in which the ω parameter, which quantifies the degree to which behavior in the pavlovian approach task is guided by MB and/or FMF learning, was fixed at a value of 1 (e.g., no FMF contribution to the action values) or at a value of 0 (e.g., no MB contribution to the action values). The model in which the ω was fixed at a value of 0 had the lowest Bayesian information criterion (BIC), indicating the FMF-only model best explained the behavior of most rats. This was consistent with the distribution of the pavlovian scores we observed (Fig. 2C), indicating that most rats in the current study exhibited high ST behaviors. The BIC for rats that had the strongest goal-tracking behavior, however, was lowest when

the ω was fixed at a value of 1, indicating that the full hybrid model is only required for some individuals. These results suggest that although the FMF-only model (e.g., $\omega = 0$) is sufficient in explaining the behavior of most rats in the current study, this is likely an artifact of the large proportion of ST, and few GT, rats in the current cohort and would not be the case with larger samples sizes consisting of more GT rats. Because the current study sought to characterize behavioral variation at an individual level, we believe that the hybrid model in which the ω is a free parameter and can vary for each individual rat is better suited to achieve this goal.

We found that some of the parameter estimates were on extreme ends of the distribution and/or boundary, likely because we were trying to optimize five parameters with a limited number of trials (~125 trials/rat). To improve model fit, we fixed four of the parameters ($\alpha, \beta, \gamma, u_{ITT}$) to the median value estimate obtained from the hybrid model and optimized only the ω parameter for each individual rat, as we have previously done (Lesaint et al., 2014). The BIC of this reduced model was lower than the FMF-only model (Table 2), and the ω parameter estimate distribution was found to be less extreme than those observed with the hybrid model (Fig. 3A). Moreover, the ω parameter estimate from the full model was correlated with the ω parameter obtained from the restricted model (Spearman's $\rho = 0.87$; $p < 0.001$; Fig. 3B), suggesting that the restricted model with only a single free parameter (e.g., ω parameter) was able to capture the individual differences observed with the full hybrid model. Our subsequent analyses involving the ω parameter were those estimates obtained using the restricted model.

To ensure that the ω parameter estimate was not being skewed by the dynamics of learning that occurs across the five pavlovian sessions, we also estimated the ω parameter using the trial-by-trial data collected in the last two pavlovian sessions (e.g., 50 trials in total). We then compared this estimate that the ω parameter obtained from trial-by-trial data collected in all the pavlovian sessions (e.g., 125 trials). The ω parameter estimates from these analyses were positively correlated with one and other (Spearman's $\rho = 0.41$; $p = 0.08$), suggesting that inclusion of earlier sessions when learning was occurring did not bias our estimate of the ω parameter. Subsequent analyses reported below were done using the ω parameter that was estimated from trial-by-trial data from all five pavlovian sessions.

Our previous simulation experiments using this reinforcement-learning model have found that as the ω parameter approaches zero, and the decision-making algorithm favors a FMF system, the prevalence of sign-tracking behaviors increases. We hypothesized, therefore, that the ω parameter would be negatively correlated with the summary pavlovian scores across rats. Indeed, the ω parameter that was estimated from the trial-by-trial data collected across the five pavlovian sessions rats completed was negatively correlated with the summary PavCA score (Spearman's $\rho = 0.89$; $p < 0.001$; Fig. 3C). These results, collectively, indicate that the restricted hybrid reinforcement-learning model can capture meaningful variation in pavlovian approach behavior.

Reward-guided behavior in the deterministic MSDM task is related to ST behaviors

Choice behavior on the deterministic MSDM task was then examined (Fig. 4A,B). The probability that rats selected the first-stage choice associated with the most frequently reinforced second-stage option increased across the 35 training sessions ($\beta = 0.012$, $p < 0.001$) and was significantly greater than that

Table 2. Goodness-of-fit measures for the full hybrid model and other variants

Parameter type	Hybrid model	FMF only ($\omega = 0$)	MB only ($\omega = 1$)	Reduced model
Free parameters	$\alpha, \beta, \gamma, \omega, u_{ITT}$	α, β, u_{ITT}	$\alpha, \beta, \gamma, u_{ITT}$	ω
Fixed parameters	NA	ω	ω	$\alpha, \beta, \gamma, u_{ITT}$
BIC	2019	1891	2147	1792

NA, Non applicable.

predicted by chance in the last five sessions that rats completed (binomial test, $p < 0.001$; Fig. 4C). Rats were more likely to repeat a first-stage choice that was subsequently rewarded than a first-stage choice that was subsequently unrewarded (Wald $\chi^2 = 113.57$, $p < 0.001$; Fig. 4D) indicating that second-stage outcomes were able to influence subsequent first-stage choices. These data, collectively, indicate that rats understood the structure of the deterministic MSDM task and, critically, that their first-stage choices influenced the subsequent availability of second-stage options.

To quantify the influence of previous trial events (e.g., correct, outcome) on first-stage choices, choice data from rats was analyzed with a logistic regression model (Fig. 4E, Table 3). The intercept was significantly greater than zero ($z = 14.92$, $p < 0.001$), indicating that rats, similar to humans, were more likely to repeat a first-stage choice regardless of previous trial events. Nevertheless, the effect of outcome was also significantly different from zero ($z = 46.56$, $p < 0.001$), indicating that rats were using previous trial outcomes (reward and absence of reward) to guide their first-stage choices.

We then examined whether individual differences in pavlovian approach behavior predicted choice behavior of the same rat in the deterministic MSDM task. The summary pavlovian score was included as a covariate in the logistic regression model and the two-way interaction between outcome and the pavlovian score examined. The summary pavlovian score \times outcome interaction was a significant predictor in the model ($z = 7.51$, $p < 0.001$; Table 3), and *post hoc* analyses indicated that the regression coefficient for outcome was significantly greater in high ST rats compared with the low ST rats ($z = 9.58$, $p < 0.001$; Fig. 4F). These data demonstrate that high ST rats were more likely to use previous trial outcomes to guide their choice behavior compared with low ST rats.

The outcome regression coefficient quantifies the degree to which both rewarded and unrewarded outcomes guide subsequent choice behavior. Differences in the outcome regression coefficient that we observed between high and low ST rats might, therefore, reflect variation in how rats use rewarded or unrewarded outcomes to guide their behavior. To independently assess the impact of rewarded and unrewarded trials on first-stage choices, we conducted a second logistic regression analysis of choice data in the deterministic MSDM task which examined the likelihood that rats would repeat the same first-stage choice following a rewarded or unrewarded outcome. The rewarded regression coefficient was positive ($\beta = 1.98 \pm 0.03$, $z = 71.59$, $p < 0.001$), indicating that rats repeated first-stage choices that resulted in reward. The unrewarded regression coefficient was also positive ($\beta = 0.33 \pm 0.02$, $z = 17.78$, $p < 0.001$) but smaller than that for rewarded regression coefficient (Wald $\chi^2 = 106$, $p < 0.001$), indicating that rats were more likely to repeat rewarded first-stage choices than unrewarded first-stage choices.

We then examined whether the summary pavlovian score interacted with the rewarded or unrewarded regression coefficients to predict first-stage choices in the deterministic MSDM

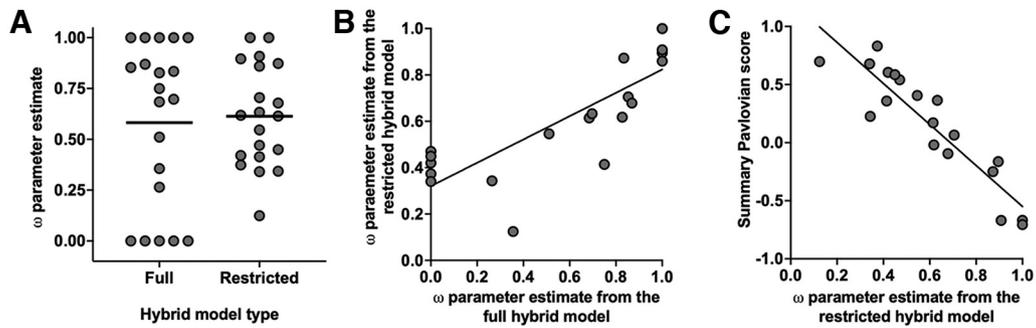


Figure 3. The hybrid reinforcement model for assessing model-based and model-free mechanisms of pavlovian learning. **A**, The ω parameter estimate in the full (left) and restricted (right) hybrid model. **B**, The relationship between the ω parameter in the full hybrid model and the ω parameter from the restricted hybrid model. **C**, The relationship between the ω parameter from the restricted hybrid model, estimated from trial-by-trial data for all five pavlovian sessions, and the summary PavCA score, measured in the last two pavlovian sessions.

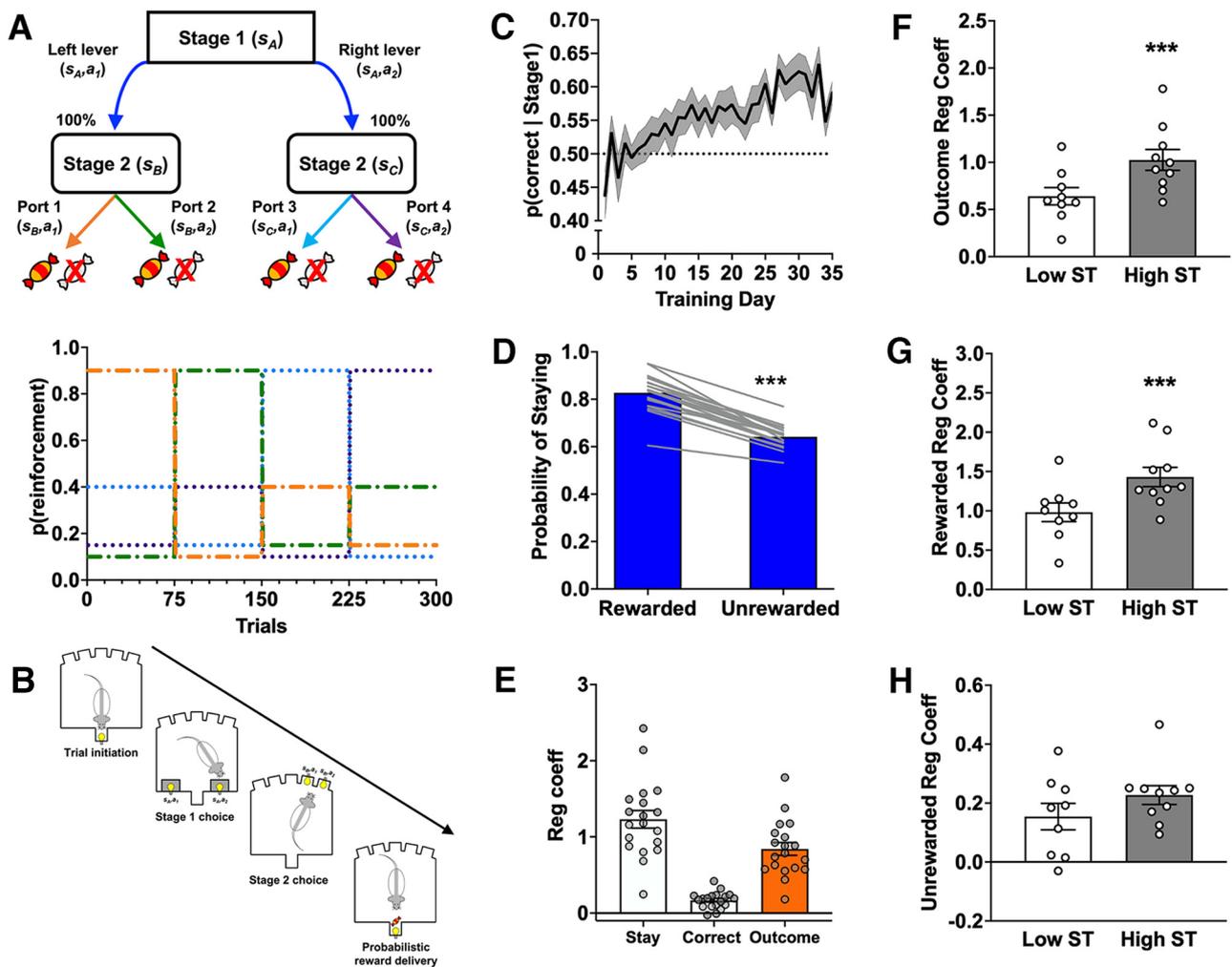


Figure 4. Decision-making in the deterministic MSDM task. **A**, Top, Rats were trained on the MSDM task in which state transitions were deterministic. Bottom, Stage 2 choices were reinforced according to an alternating block schedule of reinforcement. **B**, Schematic of single-trial events. Rats initiated trials by entering an illuminated magazine. Two levers (stage 1) located on either side of the magazine were extended into the operant box, and a single-lever response led to the illumination of two port apertures (stage 2) located on the panel opposite the levers. Entries into the illuminate apertures resulted in probabilistic delivery of reward. **C**, Probability of selecting the stage 1 option associated with the highest reinforced stage 2 options [$p(\text{correct} | \text{stage}1)$] during the first 35 d of training. The probability that choices were at chance level is represented by the dashed line. **D**, The probability that individual rats would choose the same first-stage option following a rewarded or unrewarded second-stage choice. **E**, Regression coefficients for the explanatory variables (e.g., correct, outcome) in the logistic regression model predicting the likelihood that rats would make the same first-stage choice on the current trial as they had on the previous trial in the deterministic MSDM task. Positive regression coefficients indicate a greater likelihood that the rat will repeat the same first-stage choice. **F**, The outcome regression coefficient was higher in high ST rats compared with low ST rats, indicating that second-stage outcomes were guiding first-stage choices to a greater degree in high ST rats. **G**, The likelihood that rats would repeat the same first-stage choice following a rewarded outcome was greater in high ST rats compared with low ST rats as evidenced by differences in the rewarded regression coefficient. **H**, The likelihood that rats would repeat the same first-stage choice following an unrewarded outcome did not differ between high and low ST rats. *** $p < 0.001$. Values presented are mean \pm SEM.

Table 3. Logistic regression for the deterministic MSDM task

Independent variable	Beta	Z value	p value
Intercept	1.12	14.16	<0.001
Correct	0.17	11.89	<0.001
Outcome	0.73	46.48	<0.001
Summary PavCA score	0.31	1.90	0.06
Outcome × summary PavCA score	0.27	8.54	<0.001

Table 4. Simple logistic regression for the deterministic MSDM task

Independent variable	Beta	Z value	p value
Intercept	−0.02	−0.26	0.795
Correct	−0.05	−3.88	<0.001
Rewarded trial	1.91	75.13	<0.001
Unrewarded trial	0.33	18.78	<0.001
Summary PavCA score	−0.06	−0.30	0.76
Rewarded × summary PavCA score	0.54	10.47	<0.001
Unrewarded × summary PavCA score	0.001	0.03	0.98

(Table 4). The interaction between the summary pavlovian score times rewarded regression coefficient was significant ($z = 8.93$, $p < 0.001$, $\beta = 0.51$) and *post hoc* analyses between the low and high ST groups indicated that the rewarded regression coefficient was greater in high ST rats compared with low ST rats ($z = 12.89$, $p = 0.001$; Fig. 4G). The summary pavlovian score times unrewarded interaction, however, was not significant ($z = 0.27$, $p = 0.79$, $\beta = 0.01$; Fig. 4H). To confirm these differences in outcome-specific behaviors, we compared the probability that rats would repeat a first-stage choice following a rewarded (e.g., win-stay) or unrewarded (e.g., lose-stay) outcome between high and low ST rats. The probability of repeating a first-stage choice following a rewarded outcome was greater in high ST rats compared with low ST rats (Wald $\chi^2 = 5.77$, $p = 0.02$). No differences were observed for the probability of repeating a first-stage choice following an unrewarded outcome (Wald $\chi^2 = 1.50$, $p = 0.22$). High ST rats, therefore, used rewarded outcomes to guide their first-stage choices to a greater degree than low ST rats, suggesting that these former individual differences in pavlovian incentive learning are associated with variation in reward-guided instrumental behavior.

Probabilistic MSDM task and relationship to pavlovian approach behavior

To determine whether the relationship between the summary pavlovian score and reward-guided behavior in the above deterministic version of the MSDM task was associated specifically with model-free or model-based reinforcement learning, the choice behavior of rats was assessed in the probabilistic version of the MSDM task (Fig. 5A). According to model-free theories of reinforcement learning, the probability of repeating a first-stage choice should be influenced only by the previous trial outcome, regardless of whether the state transition was common or rare (Fig. 5B, left). In contrast, model-based theories of reinforcement learning posit that the outcome at the second stage should affect the choice of the first-stage option differently based on the state transition that was experienced (Fig. 5B, right). Evidence in humans and in our previous rodent studies, however, indicates that individuals use a mixture of model-free and model-based strategies in the probabilistic MSDM task. Indeed, the probability that rats in the current study would repeat the same first-stage choice according to outcomes received (rewarded or unrewarded) and the

state transitions experienced (common or rare) during the immediately preceding trial indicated that rats were using both model-free and model-based learning to guide their choice behavior (Fig. 5C).

To quantify the influence of model-free and model-based strategies, choice data were analyzed with a logistic regression model (Daw et al., 2011; Akam et al., 2015, 2021; Groman et al., 2019a, b). The main effect of outcome, which provides an index of model-free learning, was significantly greater than zero ($z = 22.65$, $p < 0.001$; Fig. 5D, orange bar), indicating that rats were using second-stage outcomes to guide their first-stage choices. The interaction between the previous trial outcome and state transition, which provides an index of model-based learning, was also significantly greater than zero ($z = 15.38$, $p < 0.001$; Fig. 5D, purple bar). The combination of a significant main effect for outcome and a significant transition-by-outcome interaction suggests that rats were using both model-free and model-based strategies to guide their choice behavior in the probabilistic MSDM task.

We then examined whether the summary pavlovian score interacted with model-free and/or model-based learning to predict the probability of repeating the same first-stage choice in the probabilistic MSDM task (Table 5). The interaction between the summary pavlovian score and trial outcome significantly predicted choice behavior ($z = 3.16$, $p = 0.002$), but the interaction between the summary pavlovian score and the outcome-by-transition predictor did not ($z = 1.60$, $p = 0.11$). *Post hoc* comparisons between low and high ST rats indicated that the outcome regression coefficient—a measure of model-free learning—was significantly greater in high ST rats compared with low ST rats ($z = 2.67$, $p = 0.008$, $\beta = 0.09$; Fig. 5E), which was a similar effect observed in the deterministic task (Fig. 4F). The outcome-by-transition regression coefficient—a measure of model-based learning—did not differ between the low and high ST rats (Fig. 5F). These differences in the outcome regression coefficient (e.g., model-free learning) and lack of differences in the outcome-by-transition coefficient (e.g., model-based learning), collectively, indicate that high ST rats rely to a great degree on model-free learning in the MSDM task compared with low ST rats.

The greater model-free learning we observed in high ST rats may be, in part, because high ST rats acquired greater incentive value for the lever used in the pavlovian conditioning task that then biased responding in the MSDM task. We hypothesized that if this were true, then model-free behavior for the lever used in the pavlovian task might be higher than model-free behavior for the lever that was not used in the pavlovian task. To test this hypothesis, the probability that rats would repeat the same first-stage choice based on the second-stage outcomes (rewarded vs unrewarded) and state transition (common vs rare) was calculated for each lever. The difference between the probability of repeating a rewarded first-stage choice and an unrewarded first-stage choice was calculated to obtain an index of model-free learning for each lever. We compared the lever-specific index based on whether the lever in the MSDM task was in the same location as the lever used in the pavlovian task (referred to as “same”) or was in a different location as the lever used in the pavlovian task (referred to as “different”). We found that the model-free index did not differ between the levers (same lever, 0.21 ± 0.05 ; different lever, 0.28 ± 0.06 ; Wald $\chi^2 = 0.77$, $p = 0.38$). Notably, the model-free index did not differ between the levers in the high ST rats (same lever, 0.26 ± 0.05 ; different lever, 0.27 ± 0.07 ; Wald $\chi^2 = 0.02$; $p = 0.90$), suggesting that

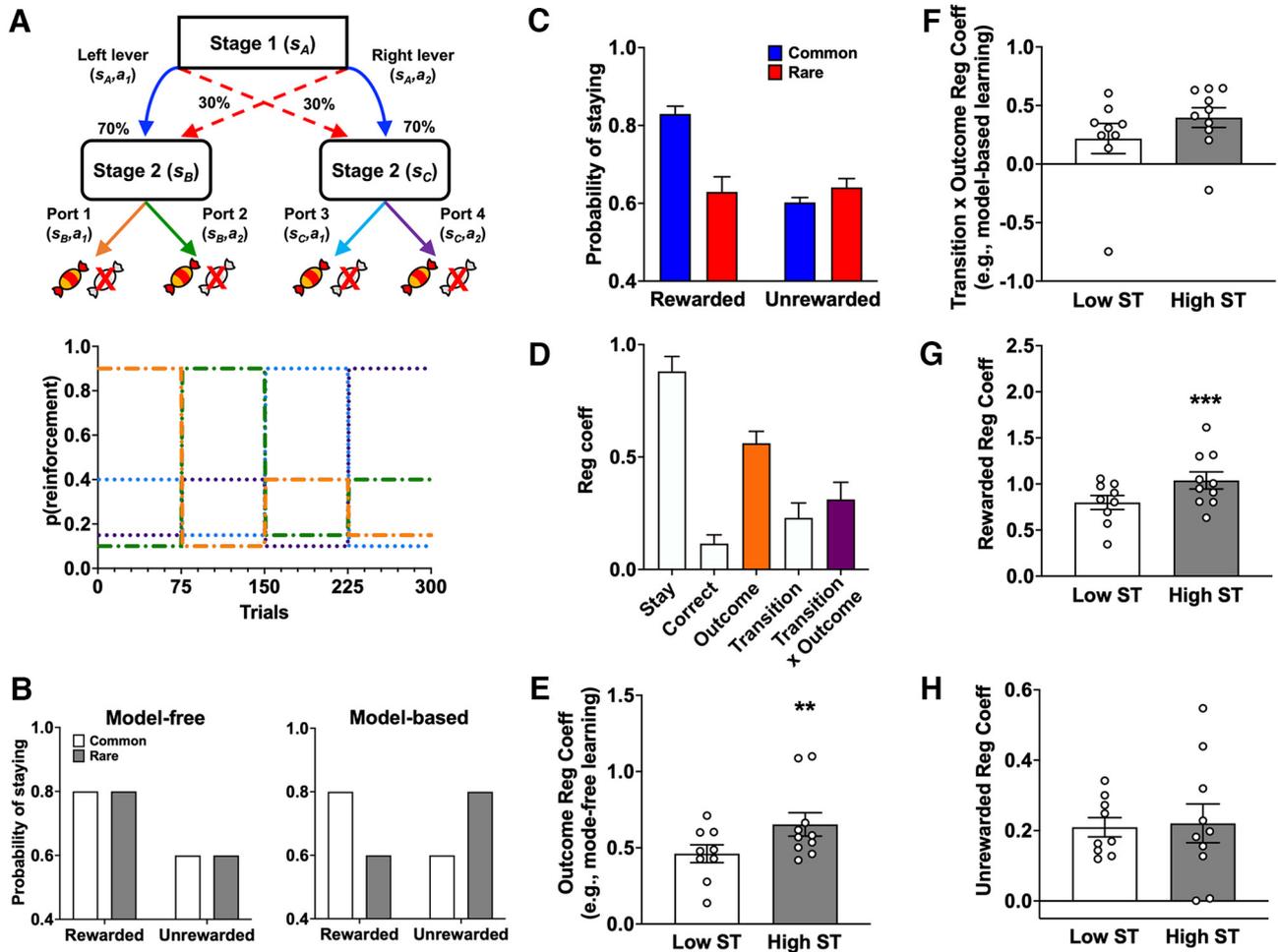


Figure 5. Decision-making in the probabilistic MSDM task. **A**, Choice behavior was assessed in the probabilistic MSDM task, which was similar in structure to the reduced MSDM, but the transition between stage 1 and stage 2 was probabilistic. **B**, Hypothetical data for a pure model-free agent (left) and a pure model-based agent (right). The probability of repeating the same first-stage option based on the previous trial outcome (rewarded vs unrewarded) and the state transition (common vs rare). **C**, The probability that rats would repeat the same first-stage option based on the previous trial outcome (rewarded vs unrewarded) and the state transition (common vs rare). **D**, Regression coefficients for explanatory variables (e.g., correct, outcome, transition, and transition-by-outcome) in the logistic regression model predicting the likelihood that rats will choose the same first-stage choice as they had on the previous trial. The outcome regression coefficient (orange bar) represents the strength of model-free learning, whereas the transition-by-outcome regression coefficient (purple bar) represents the strength of model-based learning. **E**, The outcome regression coefficient, a measure of model-free learning, was greater in high ST rats compared with low ST rats. **F**, The transition-by-outcome regression coefficient did not differ between low ST and high ST rats. **G**, The rewarded regression coefficient was greater in high ST rats compared with low ST rats. **H**, The unrewarded regression coefficient did not differ between low and high ST rats. ** $p < 0.01$, *** $p < 0.001$. Values presented are mean \pm SEM.

Table 5. Logistic regression for the probabilistic MSDM task

Independent variable	Beta	Z value	p value
Intercept	0.87	15.20	<0.001
Correct	0.09	6.60	<0.001
Outcome	0.37	19.77	<0.001
Transition	0.16	8.77	<0.001
Outcome times transition	0.26	13.80	<0.001
Summary pavlovian score	0.09	0.80	0.42
Outcome \times summary PavCA score	0.13	3.49	<0.001
Transition \times summary PavCA score	0.07	1.88	0.06
Outcome \times transition times summary PavCA score	0.05	1.41	0.16

prior experience with one of the levers in the pavlovian conditioning task did not bias high ST rats to use a model-free strategy in the MSDM task.

To determine whether the summary pavlovian score was associated with rewarded or unrewarded outcomes, choice behavior in the probabilistic MSDM task was analyzed with an alternative logistic regression model (Table 6). Similar to

Table 6. Simple logistic regression for the probabilistic MSDM task

Independent variable	Beta	Z value	p value
Intercept	0.05	0.36	0.72
Correct	-0.03	-2.46	0.01
Rewarded trial	1.45	56.55	<0.001
Unrewarded trial	0.36	20.35	<0.001
Summary pavlovian score	0.11	0.44	0.66
Rewarded trial \times summary PavCA score	0.22	4.40	<0.001
Unrewarded trial \times summary PavCA score	-0.09	-2.42	0.02

what we had observed in the deterministic MSDM task, the interaction between the summary pavlovian score and rewarded predictor was significant ($z = 4.31, p < 0.001, \beta = 0.23$), high ST rats were more likely to repeat a first-stage choice that led to a rewarded second-stage choice compared with low ST rats (Fig. 5G). We also observed a significant interaction between the summary pavlovian score and the unrewarded predictor ($z = -2.69, p = 0.007, \beta = -0.09$), but the unrewarded regression coefficient was not statistically different between low and high

ST rats (Fig. 5H). Moreover, the probability that rats would repeat a first-stage choice following a rewarded, but not unrewarded, outcome was greater in high ST rats compared with low ST rats (rewarded, Wald $\chi^2 = 4.39$, $p = 0.04$; unrewarded, Wald $\chi^2 = 0.30$, $p = 0.58$). These results, collectively, indicate that individual differences in pavlovian approach behavior are associated with variation in reward-mediated, model-free learning.

Pavlovian ST behavior is associated with reward-based, model-free updating

We found that pavlovian conditioned approach behaviors were associated with reward-mediated, model-free learning in both the deterministic and probabilistic MSDM tasks. This suggests that the model-free computations that guide pavlovian approach behaviors (e.g., FMF learning) may be related to the model-free computations that influence operant choice behavior in the MSDM task. To test this directly, we compared the regression coefficients obtained from the MSDM task in rats that either had a small ω (e.g., more model-free updating in the pavlovian conditioned approach task) or large ω (e.g., more model-based updating in the pavlovian conditioned approach task) parameter estimate (Fig. 6). We hypothesized that if the pavlovian FMF mechanisms were related to the operant-based model-free learning, then the outcome regression coefficient from the MSDM task would differ in rats with a smaller ω parameter estimate (e.g., greater FMF updating) compared with rats with a large ω parameter estimate (e.g., greater MB updating). As predicted, the outcome regression coefficient (e.g., model-free learning) was larger in rats with a smaller ω parameter compared with rats with a large ω parameter (Wald $\chi^2 = 6.22$, $p = 0.01$; Fig. 6A). These differences were specific to model-free learning, as the outcome-by-transition regression coefficient—a measure of model-based learning—did not differ as a function of the ω parameter (Wald $\chi^2 = 1.21$, $p = 0.27$; Fig. 6B). Furthermore, when we compared the rewarded and unrewarded regression coefficients between rats with either a high or low ω parameter, only the rewarded regression coefficient differed between the groups (rewarded, Wald $\chi^2 = 6.51$, $p = 0.01$; Fig. 6C; unrewarded, Wald $\chi^2 = 1.42$, $p = 0.23$; Fig. 6D). These data suggest that the model-free reinforcement-learning systems recruited during pavlovian conditioning parallel those recruited in the instrumental MSDM task.

Discussion

The current study provides new evidence that the model-free mechanisms that are used during the pavlovian conditioned approach task are related to the model-free mechanisms that guide instrumental decision-making behaviors. We report that a greater prevalence of sign-tracking behaviors in the pavlovian approach task is associated with greater model-free, but not

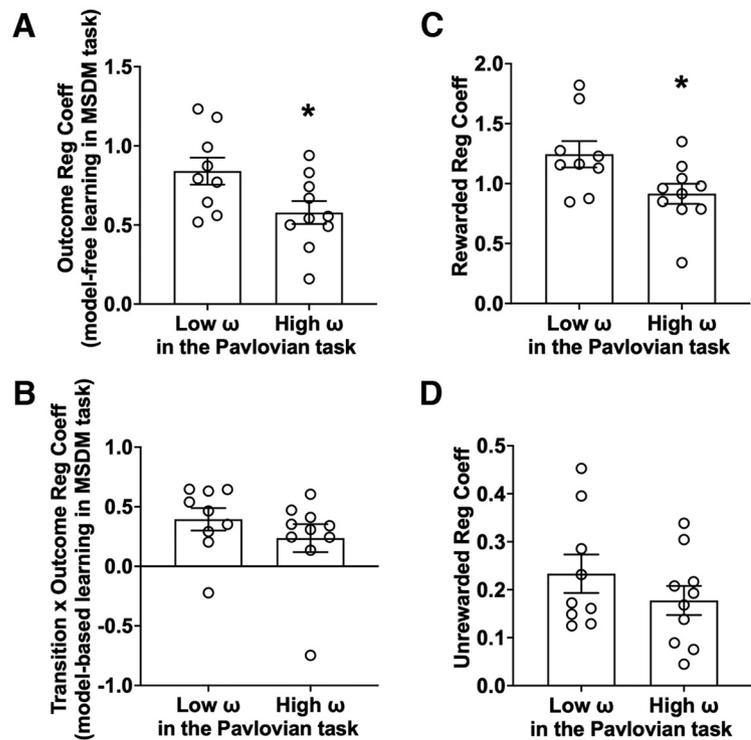


Figure 6. Model-free learning in the MSDM is related to model-free learning in the pavlovian approach task. Trial-by-trial data in the pavlovian approach task was quantified with the hybrid reinforcement learning model and the degree to which rats MB and/or FMF learning to guide their behavior quantified with the ω parameter. A median split of the ω parameter distribution was conducted, and rats classified as having a low ω parameter estimate (e.g., greater FMF updating and sign-tracking behaviors) or a high ω parameter estimate (e.g., greater MB updating and goal-tracking behaviors). **A**, The outcome regression coefficient, a measure of model-free learning in the probabilistic MSDM task, was greater in rats with a low ω parameter estimate compared with rats with a high ω parameter estimate. **B**, The transition-by-outcome regression coefficient, a measure of model-based learning in the probabilistic MSDM task, did not differ between the low and high ω parameter rats. **C**, The rewarded regression coefficient from the MSDM task was greater in rats with a low ω parameter compared with rats with a high ω parameter. **D**, The unrewarded regression coefficient from the MSDM task in rats with a low ω parameter did not differ from rats with a high ω parameter. * $p < 0.05$. Values presented are mean \pm SEM.

model-based, learning in the MSDM task. Differences in model-free updating observed in high and low ST rats were associated specifically with reward-guided behaviors; rats with higher sign-tracking behaviors were more likely to repeat a rewarded choice than rats with lower sign-tracking behaviors. No differences in choice behavior following an unrewarded outcome were observed between low and high ST rats. Our data, collectively, provide direct evidence indicating that individual differences in sign-tracking behaviors are associated with reward-based, model-free computations. These results suggest that the model-free mechanisms mediating pavlovian approach behaviors might be controlled by the same model-free computations that guide instrumental behaviors and use conserved learning systems that are known to be altered in psychiatric disorders.

Individual differences in model-free computations are conserved across instrumental and pavlovian tasks

Rats with higher sign-tracking behaviors in the pavlovian approach task were found to have greater model-free reinforcement learning in both the deterministic and probabilistic MSDM tasks. These data suggest that the mechanisms that assign and update incentive value to cues predictive of rewards might be the same as those that update representations following rewarded actions. We propose, therefore, that pavlovian and instrumental behaviors are controlled by

overlapping model-free, reinforcement-learning mechanisms. Alternatively, the related model-free measures that we quantified in the pavlovian and MSDM tasks may be driven by unique model-free mechanisms that rely on the same behavioral output. There is evidence that the neural mechanisms governing pavlovian and instrumental learning differ from each other (Bouton et al., 2021), but how these neural systems are involved in model-free computations that govern both pavlovian and instrumental learning is not fully understood. Future studies comparing how reward-mediated, model-free computations are encoded within these discrete circuits across pavlovian and instrumental environments could provide mechanistic insights into the behavioral correlations observed here.

The logistic regression analyses of choice behavior in the MSDM task indicated that rats with higher sign-tracking behaviors were more likely to repeat rewarded actions compared with rats with lower sign-tracking behaviors. This suggests that the degree of action value updating following rewards was greater in rats with higher sign-tracking behaviors and may explain why rats with greater sign-tracking behaviors are more resistant to outcome devaluation and slower to extinguish reward-predictive cues compared to GT, or lower ST, rats (Morrison et al., 2015; Nasser et al., 2015; Ahrens et al., 2016; Smedley and Smith, 2018; Fitzpatrick et al., 2019; Amaya et al., 2020; Keefer et al., 2020). For example, cached representations of cues predictive of rewards may be exaggerated in individuals with greater sign-tracking behaviors and, consequently, lead to slower adjustments in behavior when the value of the outcome changes. This is not a general impairment in extinction learning as rates of extinction of operant responses are similar between ST and GT rats (Ahrens et al., 2016; Fitzpatrick et al., 2019). Rather, previous work has proposed that strong attribution of incentive salience to reward-predictive cues may bias attention and lead to inflexible patterns of responding (Nasser et al., 2015; Ahrens et al., 2016; Keefer et al., 2020). Indeed, this may explain why sign-tracking behaviors in rats are associated with suboptimal choice behavior in a gambling task (Swintosky et al., 2021).

We did not, however, observe a relationship between pavlovian approach behaviors and model-based updating in the MSDM task. This was surprising given our previous theoretical work and the experimental work of others (Lesaint et al., 2014; Cinotti et al., 2019). The lack of association between the pavlovian summary score and model-based learning in the MSDM task is likely because we only observed a limited number of GT rats in the current sample. Specifically, only three rats in the current cohort of 20 would have been classified as GT rats (Fig. 2). This was not because the distribution of pavlovian approach behaviors in the current study was abnormal; previous studies using larger sample sizes than the current study (e.g., $N = 560$ vs $N = 20$) have observed similarly skewed distributions (Fitzpatrick et al., 2013) in food-restricted rats (Fraser and Janak, 2017). It is possible that our food restriction procedure biased rats toward a more model-free strategy in both pavlovian and instrumental environments. Future studies that use large sample sizes and manipulate hunger states to obtain behavioral measures that span the distribution of pavlovian approach behaviors may, therefore, find a relationship between goal-directed behaviors and model-based learning.

Prior experience with a particular lever in the pavlovian conditioning task did not appear to bias the behavior of rats in the MSDM task. It is possible, however, that the use of levers in both the pavlovian and operant environments had a more general influence on behavior in the MSDM task, and this influence was greater in high ST rats that attributed greater incentive salience

to the lever. Although the testing environments and outcomes (e.g., sucrose pellet vs sweetened condensed milk solution) used for the pavlovian and MSDM tasks were different from one another, randomizing the order in which animals proceeded through each of the tasks would have reduced any potential order effects that may be confounding our results. We did consider implementing a crossover design to reduce any potential order effects but believed that extensive training in the MSDM task first, compared with the limited exposure in the pavlovian conditioning task, was more likely to have an impact on behavior in the pavlovian task. A more optimal design would have used different manipulandum in the pavlovian and instrumental tasks. Nevertheless, this is a limitation of the current study design, which we will address in future experiments.

The current study was only conducted in male rats, which limits our understanding of how these pavlovian and instrumental reward-based, model-free systems interact in females. Previous studies have not reported robust differences in the prevalence of sign-tracking and/or goal-tracking behaviors between male and female rats (Pitchers et al., 2015) or model-free and model-based learning in male and female humans (Gillan et al., 2015). We would not anticipate observing different results in female rats from those reported here in male rats. Nevertheless, it is possible that the model-free mechanisms mediating pavlovian approach behaviors in females are not the same model-free computations that guide instrumental behavior. This might explain the divergent learning strategies that have been observed between male and female mice (Chen et al., 2020).

Neurobiological mechanisms

Although the neurobiological mechanisms underlying pavlovian and instrumental learning are not fully understood, dopamine neurotransmission is likely to be a point of convergence between sign-tracking behaviors and reward-guided, model-free updating. Midbrain dopamine neurons are known to encode reward-prediction errors (RPEs), which is a fundamental computation in model-free learning (Hollerman and Schultz, 1998). The results of studies using voltammetry to quantifying changes in dopamine concentration in the nucleus accumbens, a main output of midbrain dopamine neurons, have proposed that phasic dopamine signals in ST rats is how incentive salience is transferred from the outcome to cue(s) predictive of reward (e.g., lever extension; Flagel et al., 2011). These dopaminergic RPEs were not observed in goal-tracking rats, suggesting that variation in attribution of incentive salience may reflect underlying differences in dopaminergic RPEs (Derman et al., 2018; Lee et al., 2018). Indeed, antagonism of dopamine signaling in the nucleus accumbens attenuates the expression of sign-tracking behaviors (Saunders and Robinson, 2012).

Dopamine, however, has also been implicated in model-based reinforcement learning. Individual differences in [18 F]DOPA accumulation and dopamine tone in the nucleus accumbens of humans and rats, respectively, are associated with variation in model-based learning in the MSDM task (Deserno et al., 2015; Groman et al., 2019a). Dopamine may play a role in both reinforcement-learning systems. Indeed, previous studies have reported that both model-free and model-based calculations are encoded in the activity of midbrain dopamine neurons (Sadacca et al., 2017; Sharpe et al., 2017; Keiflin et al., 2019), but the influence of these dopaminergic neurons over behavior, and likely learning systems, is mediated by functionally heterogeneous circuits (Keiflin and Janak, 2015; Saunders et al., 2018). For example,

mesocortical dopaminergic projections may encode model-based computations, whereas mesostriatal/mesopallidal dopaminergic projections may encode model-free computations (Chang et al., 2015). Studies that integrate circuit-based imaging approaches with biosensor technology (e.g., dLight) to measure circuit-specific dopamine transients in behaving animals could help resolve these critical questions regarding the functional role of dopamine circuits in these learning mechanisms (Kuhn et al., 2018).

Implications for addiction

Differences in the degree to which individuals attribute incentive salience to cues predictive of reward have been hypothesized to confer vulnerability to addiction. Indeed, there is evidence that ST rats will work hard to obtain cocaine (Saunders and Robinson, 2011), show greater cue-induced reinstatement (Saunders and Robinson, 2010; Saunders et al., 2013; Everett et al., 2020), are resistant to punished drug use (Saunders et al., 2013; Pohoróalá et al., 2021), have a greater propensity for psychomotor sensitization (Flagel et al., 2008), and also display a higher preference for cocaine over food (Tunstall and Kearns, 2015) compared with GT rats. Drug self-administration in short access sessions, however, does not differ between ST and GT rats (Saunders and Robinson, 2011; Pohoróalá et al., 2021). These data suggest that drug reinforcement may be similar between ST and GT rats but that ST rats may be more susceptible or prone to developing compulsive-like behaviors following initiation of drug use.

Only a few studies have used the MSDM task to examine the role of model-free and model-based learning in addiction susceptibility. In a previous study we reported that individual differences in model-free learning in the MSDM task were predictive of methamphetamine self-administration in long-access sessions (Groman et al., 2019b). This relationship, however, was negative; rats with lower model-free learning in the MSDM task took more methamphetamine than rats with higher model-free learning. Although additional addiction-relevant behaviors were not assessed in this previous study (e.g., progressive ratio, extinction, or reinstatement), the negative relationship between model-free learning and methamphetamine self-administration is surprising given the positive relationship between model-free learning and sign-tracking behaviors we observed here. These data might suggest a dynamic role of model-free learning in the different stages of addiction susceptibility (Kawa et al., 2016). For example, greater model-free learning before drug use may protect against drug intake but render individuals more vulnerable to the detrimental effects of the drug when ingested. Indeed, ST rats are less sensitive to the acute locomotor effects of cocaine but have a greater propensity for psychomotor sensitization (Flagel et al., 2008). Future studies that assess pavlovian conditioned approach behaviors and instrumental reinforcement-learning mechanisms in the same individual before evaluating drug-taking and drug-seeking behaviors may provide a greater understanding of the biobehavioral mechanisms underlying addiction susceptibility.

Summary

The present article provides direct evidence linking incentive salience processes with reward-guided, instrumental behaviors in adult male rats. Our data suggest that pavlovian approach behaviors and choice behavior of rats in a multistage decision-making task are driven by conserved model-free reinforcement-learning mechanisms that are known to be altered in individuals with mental illness, such as addiction (Groman et al., 2022).

Future studies integrating systems-level approaches with the sophisticated behavioral and computational approaches used here will provide new insights into the biobehavioral mechanisms that are altered in individuals with mental illness.

References

- Ahrens AM, Singer BF, Fitzpatrick CJ, Morrow JD, Robinson TE (2016) Rats that sign-track are resistant to pavlovian but not instrumental extinction. *Behav Brain Res* 296:418–430.
- Akam T, Costa R, Dayan P (2015) Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. *PLoS Comput Biol* 11:e1004648.
- Akam T, Rodrigues-Vaz I, Marcelo I, Zhang X, Pereira M, Oliveira RF, Dayan P, Costa RM (2021) The anterior cingulate cortex predicts future states to mediate model-based action selection. *Neuron* 109:149–163.e7.
- Amaya KA, Stott JJ, Smith KS (2020) Sign-tracking behavior is sensitive to outcome devaluation in a devaluation context-dependent manner: implications for analyzing habitual behavior. *Learn Mem* 27:136–149.
- Boakes RA (1977) Performance on learning to associate a stimulus with positive reinforcement. In: *Operant-pavlovian interactions*. (Davis H, Hurwitz HMB, eds), pp 67–97. Hillsdale, NJ: Lawrence Erlbaum.
- Bouton ME, Maren S, McNally GP (2021) Behavioral and neurobiological mechanisms of pavlovian and instrumental extinction learning. *Physiol Rev* 101:611–681.
- Chang SE, Todd TP, Bucci DJ, Smith KS (2015) Chemogenetic manipulation of ventral pallidal neurons impairs acquisition of sign-tracking in rats. *Eur J Neurosci* 42:3105–3116.
- Chen CS, Ebitz RB, Bindas SR, Redish AD, Hayden BY, Grissom NM (2020) Divergent strategies for learning in males and females. *Curr Biol* 31:39–50.e4.
- Cinotti F, Marchand AR, Roesch MR, Girard B, Khamassi M (2019) Impacts of inter-trial interval duration on a computational model of sign-tracking vs. goal-tracking behaviour. *Psychopharmacology (Berl)* 236:2373–2388.
- Culbreth AJ, Westbrook A, Daw ND, Botvinick M, Barch DM (2016) Reduced model-based decision-making in schizophrenia. *J Abnorm Psychol* 125:777–787.
- Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69:1204–1215.
- Dayan P, Berridge KC (2014) Model-based and model-free pavlovian reward learning: reevaluation, revision, and revelation. *Cogn Affect Behav Neurosci* 14:473–492.
- Derman RC, Schneider K, Juarez S, Delamater AR (2018) Sign-tracking is an expectancy-mediated behavior that relies on prediction error mechanisms. *Learn Mem* 25:550–563.
- Deserno L, Huys QJM, Boehme R, Buchert R, Heinze H-J, Grace AA, Dolan RJ, Heinz A, Schlagenhauf F (2015) Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proc Natl Acad Sci U S A* 112:1595–1600.
- Doñamayor N, Ebrahimi C, Garbusow M, Wedemeyer F, Schlagenhauf F, Heinz A (2021) Instrumental and pavlovian mechanisms in alcohol use disorder. *Curr Addict Rep* 8:156–180.
- Everett NA, Carey HA, Cornish JL, Baracz SJ (2020) Sign tracking predicts cue-induced but not drug-primed reinstatement to methamphetamine seeking in rats: effects of oxytocin treatment. *J Psychopharmacol* 34:1271–1279.
- Fitzpatrick CJ, Gopalakrishnan S, Cogan ES, Yager LM, Meyer PJ, Lovic V, Saunders BT, Parker CC, Gonzales NM, Aryee E, Flagel SB, Palmer AA, Robinson TE, Morrow JD (2013) Variation in the form of pavlovian conditioned approach behavior among outbred male Sprague-Dawley rats from different vendors and colonies: sign-tracking vs goal-tracking. *PLoS One* 8:e75042.
- Fitzpatrick CJ, Geary T, Creeden JF, Morrow JD (2019) Sign-tracking behavior is difficult to extinguish and resistant to multiple cognitive enhancers. *Neurobiol Learn Mem* 163:107045.
- Flagel SB, Watson SJ, Akil H, Robinson TE (2008) Individual differences in the attribution of incentive salience to a reward-related cue: influence on cocaine sensitization. *Behav Brain Res* 186:48–56.
- Flagel SB, Akil H, Robinson TE (2009) Individual differences in the attribution of incentive salience to reward-related cues: implications for addiction. *Neuropharmacology* 56:139–148.

- Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, Akers CA, Clinton SM, Phillips PEM, Akil H (2011) A selective role for dopamine in stimulus-reward learning. *Nature* 469:53–757.
- Fraser KM, Janak PH (2017) Long-lasting contribution of dopamine in the nucleus accumbens core, but not dorsal lateral striatum, to sign-tracking. *Eur J Neurosci* 46:2047–2055.
- Gillan CM, Otto AR, Phelps EA, Daw ND (2015) Model-based learning protects against forming habits. *Cogn Affect Behav Neurosci* 15:523–536.
- Groman SM, Massi B, Mathias SR, Curry DW, Lee D, Taylor JR (2019a) Neurochemical and behavioral dissections of decision-making in a rodent multistage task. *J Neurosci* 39:295–306.
- Groman SM, Massi B, Mathias SR, Lee D, Taylor JR (2019b) Model-free and model-based influences in addiction-related behaviors. *Biol Psychiatry* 85:936–945.
- Groman SM, Lee D, Taylor JR (2021) Unlocking the reinforcement-learning circuits of the orbitofrontal cortex. *Behav Neurosci* 135:120–128.
- Groman SM, Thompson SL, Lee D, Taylor JR (2022) Reinforcement learning detuned in addiction: integrative and translational approaches. *Trends Neurosci* 45:96–105.
- Hammersley R (1992) Cue exposure and learning theory. *Addict Behav* 17:297–300.
- Hearst E, Jenkins HM (1974) Sign-tracking: the stimulus-reinforcer relation and directed action. Austin TX: Psychonomic Society.
- Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci* 1:304–309.
- Huys QJM, Tobler PN, Hasler G, Flagel SB (2014) The role of learning-related dopamine signals in addiction vulnerability 3. *Prog Brain Res* 211:31–77.
- Kawa AB, Bentzley BS, Robinson TE (2016) Less is more: prolonged intermittent access cocaine self-administration produces incentive-sensitization and addiction-like behavior. *Psychopharmacology (Berl)* 233:3587–3602.
- Keefer SE, Bacharach SZ, Kochli DE, Chabot JM, Calu DJ (2020) Effects of limited and extended pavlovian training on devaluation sensitivity of sign- and goal-tracking rats. *Front Behav Neurosci* 14:3.
- Keiflin R, Janak PH (2015) Dopamine prediction errors in reward learning and addiction: from theory to neural circuitry. *Neuron* 88:247–263.
- Keiflin R, Pribut HJ, Shah NB, Janak PH (2019) Ventral tegmental dopamine neurons participate in reward identity predictions. *Curr Biol* 29:93–103.e3.
- Kuhn BN, Campus P, Flagel SB (2018) The neurobiological mechanisms underlying sign-tracking behavior. In: Sign-tracking and drug addiction. (Tomie A, Morrow J, eds). Ann Arbor: Michigan Publishing, University of Michigan Library.
- Lee B, Gentry RN, Bissonette GB, Herman RJ, Mallon JJ, Bryden DW, Calu DJ, Schoenbaum G, Coutureau E, Marchand AR, Khamassi M, Roesch MR (2018) Manipulating the revision of reward value during the inter-trial interval increases sign tracking and dopamine release. *PLOS Biol* 16:e2004015.
- Lesaint F, Sigaud O, Flagel SB, Robinson TE, Khamassi M (2014) Modelling individual differences in the form of pavlovian conditioned approach responses: a dual learning systems approach with factored representations. *PLoS Comput Biol* 10:e1003466.
- Meyer PJ, Lovic V, Saunders BT, Yager LM, Flagel SB, Morrow JD, Robinson TE (2012) Quantifying individual variation in the propensity to attribute incentive salience to reward cues. *PLoS One* 7:e38987.
- Miller KJ, Botvinick MM, Brody CD (2017) Dorsal hippocampus contributes to model-based planning. *Nat Neurosci* 20:1269–1276.
- Morrison SE, Bamkole MA, Nicola SM (2015) Sign tracking, but not goal tracking, is resistant to outcome devaluation. *Front Neurosci* 9:468.
- Nasser HM, Chen YW, Fiscella K, Calu DJ (2015) Individual variability in behavioral flexibility predicts sign-tracking tendency. *Front Behav Neurosci* 9:289.
- Nasser HM, Calu DJ, Schoenbaum G, Sharpe MJ (2017) The dopamine prediction error: contributions to associative models of reward learning. *Front Psychol* 8:244.
- Pitchers KK, Flagel SB, O'Donnell EG, Woods LCS, Sarter M, Robinson TE (2015) Individual variation in the propensity to attribute incentive salience to a food cue: influence of sex. *Behav Brain Res* 278:462–469.
- Pohoróalá V, Enkel T, Bartsch D, Spanagel R, Bernardi RE (2021) Sign- and goal-tracking score does not correlate with addiction-like behavior following prolonged cocaine self-administration. *Psychopharmacology (Berl)* 238:2335–2346.
- Robinson MJF, Berridge KC (2013) Instant transformation of learned repulsion into motivational “wanting”. *Curr Biol* 23:282–289.
- Robinson TE, Flagel SB (2009) Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences. *Biol Psychiatry* 65:869–873.
- Rode AN, Moghaddam B, Morrison SE (2020) Increased goal tracking in adolescent rats is goal-directed and not habit-like. *Front Behav Neurosci* 13:291.
- Sadacca BF, Wikenheiser AM, Schoenbaum G (2017) Toward a theoretical role for tonic norepinephrine in the orbitofrontal cortex in facilitating flexible learning. *Neuroscience* 345:124–129.
- Saunders BT, Robinson TE (2010) A cocaine cue acts as an incentive stimulus in some but not others: implications for addiction. *Biol Psychiatry* 67:730–736.
- Saunders BT, Robinson TE (2011) Individual variation in the motivational properties of cocaine. *Neuropsychopharmacology* 36:1668–1676.
- Saunders BT, Robinson TE (2012) The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses. *Eur J Neurosci* 36:2521–2532.
- Saunders BT, Robinson TE (2013) Individual variation in resisting temptation: implications for addiction. *Neurosci Biobehav Rev* 37:1955–1975.
- Saunders BT, Yager LM, Robinson TE (2013) Cue-evoked cocaine “craving”: role of dopamine in the accumbens core. *J Neurosci* 33:13989–14000.
- Saunders BT, Richard JM, Margolis EB, Janak PH (2018) Dopamine neurons create pavlovian conditioned stimuli with circuit-defined motivational properties. *Nat Neurosci* 21:1072–1083.
- Sebold M, Schad DJ, Nebe S, Garbusow M, Jünger E, Kroemer NB, Kathmann N, Zimmermann US, Smolka MN, Rapp MA, Heinz A, Huys QJM (2016) Don't think, just feel the music: individuals with strong pavlovian-to-instrumental transfer effects rely less on model-based reinforcement learning. *J Cogn Neurosci* 28:985–995.
- Sharpe MJ, Chang CY, Liu MA, Batchelor HM, Mueller LE, Jones JL, Niv Y, Schoenbaum G (2017) Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat Neurosci* 20:735–742.
- Smedley EB, Smith KS (2018) Evidence for a shared representation of sequential cues that engage sign-tracking. *Behav Processes* 157:489–494.
- Swintosky M, Brennan JT, Koziel C, Paulus JP, Morrison SE (2021) Sign tracking predicts suboptimal behavior in a rodent gambling task. *Psychopharmacology (Berl)* 238:2645–2660.
- Tunstall BJ, Kearns DN (2015) Sign-tracking predicts increased choice of cocaine over food in rats. *Behav Brain Res* 281:222–228.
- Wang F, Schoenbaum G, Kahnt T (2020) Interactions between human orbitofrontal cortex and hippocampus support model-based inference. *PLoS Biol* 18:e3000578.