Sex differences in learning from exploration

Cathy S. Chen[1], Evan Knep[1], Autumn Han[1], R. Becket Ebitz[+2], Nicola M. Grissom[+1]

1 Department of Psychology, University of Minnesota, Minneapolis MN 55455
2 Department of Neurosciences, University of Montreal, Quebec, Canada

+to whom correspondence should be addressed:
Nicola Grissom
Department of Psychology
University of Minnesota
75 East River Rd
Minneapolis, MN 55455
ngrissom@umn.edu

Becket Ebitz,
Department of Neuroscience,
Faculty of Medicine
Université de Montréal
2960, chemin de la Tour, local 111
Montréal, QC H3T 1J4 Canada
becket@ebitzlab.com

Abstract

Sex-based modulation of cognitive processes could set the stage for individual differences in vulnerability to neuropsychiatric disorders. While value-based decision making processes in particular have been proposed to be influenced by sex differences, the overall correct performance in decision making tasks often show variable or minimal differences across sexes. Computational tools allow us to uncover latent variables that define different decision making approaches, even in animals with similar correct performance. Here, we quantify sex differences in mice in the latent variables underlying behavior in a classic value-based decision making task: a restless 2-armed bandit. While male and female mice had similar accuracy, they achieved this performance via different patterns of exploration. Male mice tended to make more exploratory choices overall, largely because they appeared to get "stuck" in exploration once they had started. Female mice tended to explore less, but learned more quickly during exploration. Together, these results suggest that sex exerts stronger influences on decision making during periods of learning and exploration than during stable choices. Exploration during decision making is altered in people diagnosed addictions, depression, and neurodevelopmental disabilities, pinpointing the neural mechanisms of exploration as a highly translational avenue for conferring sex-modulated vulnerability to neuropsychiatric diagnoses.

Keywords: *sex differences, exploration, exploration-exploitation tradeoff, reinforcement learning, decision making, animal models*

Introduction

Essentially every neuropsychiatric condition shows sex and/or gender biases in risk, presentation, etiology, and prognosis (Green et al., 2019; Grissom and Reyes, 2018; Shansky, 2019). This raises the possibility that sex-modulated biological mechanisms could modulate cognitive processes that confer vulnerability and/or resilience to mental health challenges. However, sex differences in cognitive task performance can be difficult to detect and even more variable than would be expected given the non-dichotomous, overlapping impacts of sex mechanisms on cognition (Maney, 2016). An underrecognized source of variability in cognitive tasks is that there can be multiple ways to achieve the same level of performance on the primary dependent variables used to assess these tasks, such as “number of correct responses”. This means that equivalent levels of performance could mask individual differences in how males and females are solving the same problem. Indeed, we have recently shown that examining the latent strategies underlying task performance --rather than differences in final performance--can reveal that individual males and females can take very different strategic paths to the learning of action-outcome associations (Chen et al., 2020). Here, we applied computational tools to characterize sex differences in the latent variables underlying behavior to understand sex differences in a key cognitive process regulating reward-guided behaviors: balancing exploration and exploitation.

In an uncertain world, we must balance two goals: exploiting rewarding options when they are available, but also exploring alternatives that could be more rewarding or provide new information about the world. Too little exploration makes behavior inflexible and perseverative. Too much makes it impossible to sustain rewarding behaviors. Exploration is dysregulated in numerous neuropsychiatric disorders (Addicott et al., 2017; Wilson et al., 2021), many of which are also sex-biased (Green et al., 2019; Grissom and Reyes, 2018; Shansky, 2019). This suggests that sex differences in exploration and exploitation could contribute to sex-linked vulnerability to these conditions, though we do not yet understand how exploration and exploitation differ with sex. Because exploration is a major source of errors in task performance more broadly (Ebitz et al., 2019; Pisupati et al., 2019), sex-differences in exploration may influence behavior in complex ways in tasks which are not intended to directly measure this tendency and could contribute to performance differences and variability seen across tasks and species (Grissom and Reyes, 2018; van den Bos et al., 2013).

To examine whether there are sex differences in exploration, we trained male and female mice on a classic explore/exploit task, a spatial restless two-armed bandit (Ebitz et al., 2018). Males showed higher levels of exploration than females throughout the task. This was because males were more likely to get "stuck" in extended periods of exploration before committing to a favored choice. On the other hand, females showed elevated reward learning specifically during bouts of exploration, making exploratory trials more informative, which allowed them to start exploiting a favored choice earlier than males. Together, these results demonstrate that while the overall performance was similar, males and females exhibited different patterns of exploration while interacting with the same uncertain environment.

Results

Age-matched male and female wildtype mice (n = 32, 16 per sex, strain B6129SF1/J) were trained to perform a restless two-armed spatial bandit task in touch-screen operant chambers (**Figure 1A**). Animals were presented with two physically identical targets (squares) on the left and right of the screen each trial and indicated their choices by nose poking at one of two target locations. Each location offered some probability of reward, which changed slowly and randomly across trials, and independently across targets. The dynamic reward contingencies encouraged the animals to constantly balance exploration and exploitation. The animals had to exploit a good option when it is found, but also occasionally explore the other option, whose drifting values meant that it could become better at any time. Mice performed 2 repetitions of 4 consecutive sessions of the restless bandit task, measuring 8 sessions total. Each session consisted of 300 trials.

To examine task performance, we first calculated the average probability of reward obtained in males and females. Because rewards schedules were stochastic, sessions could differ slightly in the amount of reward that was available. We therefore compared performance against the average probability of reward if choosing randomly within each session. Regardless of sex, mice were able to earn reward more frequently than chance (**Figure 1B,** two-way ANOVA, $F(1, 60) = 228.9$, $p < 0.0001$). There was no significant sex difference in the probability of rewards acquired above chance (main effect of sex, $F(1, 30) = 0.05$, $p = 0.83$).

Similar levels of accuracy do not require or imply a similar approach to the task. Our previous study suggested that males and females could achieve similar learning performance via divergent decision making strategies (Chen et al., 2020). However, different strategies might take different amounts of time to execute (Chen et al., 2020; Filipowicz et al., 2019; Kool et al., 2010; Kurdi et al., 2019). Therefore, if males and females had adopted strategies here, then we might expect response time to systematically differ between males and females, despite the similarities in learning performance. Indeed, females responded significantly faster than did males (**Figure 1C**, main effect of sex, $t(30) = 3.52$, $p = 0.0014$), suggesting that decision making computations may differ across sexes and, if so, that the strategies that tended to be used by females were faster than those used by males.

A Hidden Markov model (HMM) identifies distinct features of exploratory and exploitative choices in mice.

Despite similar performance, response time differences suggested that males and females employed different strategies in this task. One possible difference was sex differences in the level of exploration. Prior research has shown that exploratory choices take longer than exploitative choices (Ebitz et al., 2019, 2018). Therefore, perhaps males took longer to make a choice because a greater proportion of their choices were exploratory. To test this hypothesis, we used a hidden Markov model (HMM) to determine whether choices were exploratory or exploitative (Ebitz et al., 2018).

The HMM allows us to make statistical inferences about the probability that each choice was due to exploration or exploitation via modeling these as the latent goal states underlying choice (see Methods). Because this approach to inferring exploration is agnostic to the generative computations and depends only on the temporal statistics of choices (Ebitz et al., 2020, 2019, 2018), it is particularly ideal for circumstances like this one, where we suspect that the generative computations may differ across groups.

The HMM method has never before been used to quantify exploration in mice, so we first asked whether it was appropriate here. The HMM method works because *sequences* of exploratory decisions look very different from exploitative ones, at least in reinforcement learning agents and primates (Ebitz et al., 2018). When agents exploit, they repeatedly sample

the same option, switching only very rarely. However, because exploration requires investigatory samples, runs of exploratory choices tend not to repeat the same option. They tend to switch far more frequently, closer to what we would expect from random samples from the environment (Ebitz et al., 2018). Therefore, if mice were alternating between exploration and exploitation in this task, we would expect to see evidence of two distinct patterns of switching in their behavior. Indeed, choice run durations (i.e. the distribution of inter-switch intervals) were parsimoniously described as a mixture of two different patterns: one regime where choices switched quickly (mean switching time = 1.7 trials, compared to random choices at 2 trials; 80% of choice runs) and one regime where they changed slowly (mean switching time = 6.8 trials; **Supplemental Figure 1A**). Thus, mice had evidence of fast-switching (putatively explore) and slow-switching (putatively exploit) regimes in their behavior. Note that explore-labeled choices are more likely to also be switch choices, but not all explore-labeled choices are switches, and not all exploit-labeled choices are stay decisions.

Next, we used a simple 2-parameter HMM to infer the likelihood that each choice was part of the exploratory regime, or the exploitative one (see Methods). To evaluate the face validity of the HMM labels, we asked whether HMM-labeled exploratory choices matched the normative definition of the term. First, by definition, exploration is a pattern of non-reward-maximizing choices whose purpose is learning about rewards. This means exploratory choices should be (1) orthogonal to reward value, and (2) exhibit enhanced reward learning.

Explore-labeled choices were non-reward-maximizing: they were orthogonal to reward value (**Supplemental Figure 1B**; the average value of choices chosen during exploration was not different from chance; one sample t-test, $t(10) = 0.16$, $p = 0.87$). Reward learning was also elevated during exploration. During HMM-labeled exploratory states, the outcome of choices had more influence on the subsequent decision - animals were more likely to stay with the same choice if rewarded and switch if not rewarded (**Supplemental Figure 1C**, two-way RM ANOVA, interaction term, $F(1,31) = 51.2$, $p < 0.001$).

Finally, we asked whether the labels inferred by the HMM produced differences in behavioral metrics other than choices. We computed average choice response time separately for exploratory and exploitative states. The results suggested that choice response time was significantly longer during exploration than exploitation (**Supplemental Figure 1D**, t-test, t(31)

= 3.66, p = 0.0009), which is consistent with previous findings that exploration slows down decision making (Ebitz et al., 2018). To eliminate the possibility that exploration was merely disengagement from the task, we examined average reward retrieval time during exploratory and exploitative states. There was no significant difference in reward retrieval time between two states (**Supplemental Figure 1E**, t-test, t(31) = 0.05, p = 0.95), suggesting that animals were not more disengaged from the task during exploration than exploitation. They were only slower in making a decision. Together, these results demonstrated that HMM-labeled exploration was meaningful, non-reward-maximizing, and accompanied by enhanced reward learning, matching the normative definition of exploration.

Males made more exploratory choices than females, because they explored for longer periods of time once they started.

With more confidence in the validity of HMM-inferred states, we found that males, on average, were more likely to be in the exploratory regime than the exploitative one, with 72.9% ± 11.5% STD of trials labeled as exploratory (**Figure 1D**). Females explored much less with only 55.4% ± 20.4% STD of trials being exploratory (**Figure 1E**, t-test, t(30) = 2.98, p = 0.0056; 95% CI for the difference between the sexes = [5.5%, 29.3%]). As groups, males and females were reasonably, but not perfectly discriminable in terms of the proportion of exploratory choices (**Figure 1F,** receiver operating characteristic analysis, AUC = 0.76 ± 0.09, 95% CI for the difference = [0.59, 0.93] , p = 0.013). These differences were largely driven by the greater male tendency to keep exploring once they started. Males repeated exploration 92.1% (± 3.4% STD) of the time, while females stopped exploring and committed to a choice more quickly, repeating exploration only 83.1% (± 16.8% STD) of the time (t-test, t(30) = 2.09, p = 0.045, 95% CI for the difference = [0.2%, 17.8%]). There were no significant differences in the other parameter of the HMM (probability of repeating exploitation: males = 83.5 ± 3.7%; females = 79.7 ± 22.1%; t(30) =0.69 , p = 0.5).

Although sex differences in model parameters were modest, analyzing the full dynamics of the fitted HMMs again supported a robust sex difference in the tendency to explore (see Methods). In models fit to males, exploration was a deeper, more “sticky” behavioral state (**Figure 1G**, stationary probability of exploration = 68.0 ± 8.5% STD), compared to models fit to females, where exploration and exploitation were more closely matched (54.4 ± 18.4% STD;

different from males: $t(30) = 2.68$, $p = 0.012$, 95% CI for the difference = [3.2%, 23.9%]). This suggests that males were more likely to get "stuck" in an extended exploratory period, requiring more energy to escape from exploring and start exploiting a good choice.

Multiple variables in reinforcement learning models may be the cause of increased exploration.

The results from the HMM analyses suggest that males were, on average, more exploratory than females, and not because they were more likely to initiate exploration, but because they were more likely to become "stuck" in exploration. This suggests that there were sex differences in the animals' approach to this task. However, a crucial question remained unanswered: what computational differences made the males more exploratory? To address this question, we turned to reinforcement learning models to look for individual variability in latent cognitive parameters that could influence exploration and exploitation (Daw et al., 2006; Ishii et al., 2002; Jepma and Nieuwenhuis, 2011; Pearson et al., 2009).

We began by considering a simple, 2-parameter RL model, with one learning rate parameter (α) and one parameter for decision noise (inverse temperature β). Traditionally, only the latter parameter is thought to be related to exploration, so most previous studies of exploration have focused exclusively on this inverse temperature parameter (Beeler et al., 2010; Cinotti et al., 2019). However, we realized that exploration in an RL model is a function of both the difference in subjective values and the decision-noise (**Figure 2A**). This implied that there could be multiple ways to increase exploration even in this simple 2-parameter RL model.

Indeed, changing the inverse temperature of the Softmax function influences the probability of exploration - the larger value of β, the lower the decision noise, thus lower probability of exploration. Consistent with this prior work, we found that changing beta robustly changed the probability of exploration in simulations from the model (**Figure 2B**; GLM, main effect of inverse temperature, $\beta 1 = -0.11$, $p < 0.0001$). The second latent parameter, the learning rate, could also mediate the level of exploration by changing the value difference between options. The learning rate dictates the rate of value updating from each outcome. For example, when an agent receives two consecutive rewards (**Figure 2A** (bottom)), a higher learning rate expands the cumulative value difference between the rewarded chosen option and the unchosen

option, increasing the probability of exploiting the higher value option. The result of computer simulation confirmed this theory that learning rate also influences the probability of exploration (**Figure 2B**; GLM, main effect of learning rate, $\beta 2 = -0.10$, $p < 0.0001$). Moreover, there were some values for alpha at which changing decision noise had no effect on exploration whatsoever - alpha and beta interact to influence exploration (**Figure 2C**; GLM, interaction term, $\beta 3 = -0.38$, $p < 0.0001$). This was because when the learning rate is very low, exploration does not provide as much information about the value of the options, thus it takes longer to learn about which option is better. As a result, exploration is prolonged (duration of an exploration periods) and increased (the number of exploration bouts). Together, these results suggest that sex-differences in exploration (and perhaps individual differences in exploration more broadly) could be due to multiple different changes in the computations underlying behavior.

Females had higher learning rates than males, and showed evidence of meta-learning.

In order to understand the computational basis for the sex difference in exploration, we fit RL models in order to consider a variety of latent variables that could, either alone or in combination, influence animals' exploratory behavior. Our ability to make inferences about changes in model parameters is highly sensitive to the correct specification of the model. Therefore, we first compared the fit of six reinforcement learning (RL) models that could potentially characterize animals' choice behaviors : (1) a "random" model with some overall bias for one choice over the other, (2) a "noisy win stay lose shift" model that assumes a win stay lose shift policy with some level of randomness, (3) a two-parameter "RL" model with a consistent learning rate and some inverse temperature that captures decision noise, (4) a four-parameter "RLCK" model that captures both value-based and value-independent decision with separate parameters for learning rate, decision noise, choice bias, and choice stickiness, (5) a five-parameter "RLCKγ" model that incorporates separate learning rate for rewarded and unrewarded outcomes on top of the "RLCK" model, (6) a five-parameter "RLCKη" model that adds a parameter that tunes the weight between value-based and choice-based decision to the "RLCK" model (see methods, **Figure 3A**).

Although model fitting was slightly different across sexes, in both males and females, the "RLCK" model, four-parameter model with value and choice kernel updating policies, best characterized animals' choice behaviors in this task among all six models (**Figure 3B**). To

confirm that this model was able to capture animals' choice regardless of their learning performance, we used the parameters optimized by the model to simulate choices and compared with the actual choices of animals. We plotted the simulated choices and actual choices against the matching law (Poling et al., 2011), which dictates that the probability of choice is proportional to the probability of reward. The result supported that this four-parameter model was able to characterize animals' choice behaviors regardless of whether the animal was following the matching law (**Supplemental Figure 2**). This suggests that the best fit model, the four-parameter reinforcement learning - choice kernel (RLCK) model, captured both value-dependent and value-independent choices.

The RLCK model had 4 parameters, which we then compared across sexes. We found that females had significantly higher learning rate (α) than males (**Figure 3C**, $t(30) = 2.40$, $p = 0.02$) but there was no significant difference across sexes in other parameters (β: $t(30) = 1.44$, $p = 0.16$; α_c: $t(30) = 1.40$, $p = 0.17$; β_c: $t(30) = 1.73$, $p = 0.09$). To examine whether the higher learning rate in females was driven by a few individuals with extremely high learning rates, we plotted the distribution and calculated the separability of learning rates of two sexes. As groups, males and females were reasonably discriminable in terms of the learning rate (**Figure 3D**, receiver operating characteristic analysis, AUC = 0.72 ± 0.09, 95% CI for the difference = [0.54, 0.90], $p = 0.035$). These results suggested that the difference in the level of exploration between males and females was not due to differences in decision noise, but instead due to differences in learning rate.

Furthermore, when compared learning rate across sessions within sex, females, but not males, showed increased learning rate over experience with task (**Figure 3E**, repeated measures ANOVA, female: main effect of time, $F(2.26,33.97) = 5.27$, $p = 0.008$; male: main effect of time, $F(2.5,37.52) = 0.23$, $p = 0.84$). Females not only learned more quickly but they also "learned to learn", while males did not. This points to sex differences in meta-learning, with females learning not only the immediate choice-outcome association but also the overall dynamic reward structure of the environment (Doya, 2002).

Females learned more during exploratory choices than males.

The results of HMM model and RL models revealed significant sex differences in exploration, paralleled by sex differences in rate of learning. What remains unclear is how sex, explore-exploit states, and reward outcomes all interact together to influence the animals' choices. Therefore, we conducted a 4-way repeated measures ANOVA to examine how (1) positive and negative outcomes, (2) explore-exploit states, (3) sex, and (4) subject identity (nested in sex) all came together to influence choice: whether animals would repeat their last choice (stay) or try a different option (switch; **Table 1**). This 4-way repeated measure ANOVA allowed us to understand the main effect of sex, state, and outcome, as well as all pair and triplet-wise interaction effects, on how animals learned from previous rewards. Note that in previous analyses, we used subject averaged data but since subject average (16 subjects each sex) is underpowered to detect a 3-way interaction effect, we used session averaged data to increase the power to detect any effects across sex, state and outcome.

The results revealed an expected significant main effect of outcome on stay-switch decisions (main effect of outcome, $p < 0.00001$). This effect was driven by the tendency of animals to repeat the previous choice (i.e.: not switching) after obtaining a reward (post hoc t-test compared to chance at 0.50, mean = 0.75, 95% CI = [0.74, 0.77], $t(255)=34.33$, $p<0.0001$) and a much smaller tendency switch more often than chance after reward omission (mean = 0.52, 95%CI = [0.50, 0.55], post hoc t-test, $t(255) = 2.05$, $p=0.04$). The tendency to switch or stay also differed by sex, with females more likely to repeat a previous choice and males more likely to switch (main effect of sex, $p < 0.00001$; post-hoc t-test on p(switch): $t(254) = 4.12$, $p < 0.0001$). There was also a significant interaction effect between sex and outcome (sex X reward interaction, $p < 0.00001$). To understand how reward and reward omission differentially affect choice across sexes, we conducted post-hoc win-stay lose-shift analyses. We found that female mice displayed more win-stay behaviors, indicating that they were more likely than the males to repeat behaviors that produced reward on the previous trial (**Figure 4A**, sex X reward interaction, $p < 0.000001$, **Table 1**; post hoc t-test: $t(254) = 5.53$, $p < 0.0001$). As groups, males and females were reasonably, but not perfectly discriminable in terms of the proportion of win stay choices (**Figure 4B,** receiver operating characteristic analysis, AUC = 0.74 ± 0.09, 95% CI for the difference = [0.56, 0.92], $p = 0.0195$). In contrast, male mice tended to shift even when the previous choice was rewarded.

There was no significant sex difference in learning from losses (**Figure 4C**, $t(254) = 1.40$, $p = 0.16$), but this did not mean that sex differences in learning were solely due to sex differences in learning from wins. There were at least two ways that we could observe an equivalent tendency to lose-shift across sexes in this task. One possibility is that females only learn more from positive outcomes, but not negative ones. However, the other possibility is that this lack of a difference in lose-shift behaviors between males and females was an artifact of the tendency of males to explore more frequently. Since males spend more time exploring (**Figure 1E**) and learning from both wins and losses is enhanced during exploration (**Supplemental Figure 1D**), males could lose-shift less frequently than females during both exploration and exploitation, yet still lose-shift exactly as much as females because a greater proportion of their choices were exploratory.

To dissociate these possibilities, we next examined the effects of exploration and exploitation. The ANOVA revealed a significant main effect of state, resonating with the result of HMM model validation that animals were more likely to switch during exploratory state than during exploitative state (**Table 1**, main effect of state, $p < 0.00001$; sex X state interaction, $p = 0.0667$; **Supplemental Figure 1C**). Critically, there was also a significant 3-way interaction between sex, explore/exploit state, and reward (**Figure 4E**, **F**, **Table 1**, sex X reward X state interaction, $p = 0.0438$). This could imply sex-linked differences in reward learning across exploration and exploitation. To determine if this was true, we separated out probability of lose-shift according to whether it happened during exploration or exploitation, as inferred from the Hidden Markov model. Males switched less after losses than females within exploratory states (**Figure 4D**; post hoc t-test on session averages: sex difference within exploration: $p < 0.001$, $t(251) = 3.39$), though there was no significant sex difference within exploitation ($p = 0.06$, $t(243) = 1.87$; note that differing degrees of freedom are due to the fact that exploitation was not observed in some sessions for some animals). This supported the second hypothesis that males lose-shift less than females both when exploring and when exploiting, but that there was no difference in lose-shift overall because males spent more time in a state in which both win-stay and lose-shift choices occur more frequently (**Supplemental Figure 1C**). We also found that the increased tendency to win-stay that we observed in females was driven by the explore choices (**Figure 4E**; post-hoc t-test: $p = 0.015$, $t(251) = 2.55$). There was no significant difference in win-stay between males and females during exploit choices (post-hoc t-test: $p = 0.09$, $t(244) =$

1.68). Together these results suggest that females were better explorers (i.e.: they had increased reward learning during exploration), whereas males learned slower during exploration but compensated this learning disadvantage by exploring more frequently.

These effects were not driven by idiosyncratic strategic differences between the sexes (e.g., shifting only after two losses). We used a model-free approach to quantify the extent to which behavior was structured without making strong assumptions about what form this structure might take. We calculated conditioned mutual information for all sessions across sexes (Leao et al., 2004; Wyner, 1978), to examine how choice behavior was influenced by information of past choice history, given the immediate outcome. The result suggested that mutual information was higher in females than males, suggesting that females were using more information from the past choice and outcome to make their current decision. (**Figure 5G**, $t(30) = 2.65$, $p = 0.013$). As groups, males and females were reasonably discriminable in terms of mutual information (**Figure 5H**, receiver operating characteristic analysis, AUC = 0.74 ± 0.09, 95% CI for the difference = [0.57, 0.92], $p = 0.0195$). Together, these results reinforced our conclusion that females were learning more: utilizing more information from the past trial to make current choices.

Discussion

Sex mechanisms biasing the preferred approaches taken during cognitive tasks are potentially significant contributors to task performance. We used a combination of computational modeling approaches to characterize sex differences in a canonical explore/exploit task. While males and females had similar performance, they used different latent explore-exploit strategies to learn about the dynamic environment. Males explored more than females and were more likely to “get stuck” in an extended period of exploration before committing to a favored choice. In contrast, females were more likely to show elevated learning during the exploratory period, making exploration more efficient and allowing them to start exploiting a favored choice earlier. Furthermore, the learning rate increased over days in females but stayed stable in males. Such meta-learning in females permitted learning about the current task (which option provides the best reward outcome), as well as the structure of the task (the reward probability of choices changes over time). This allowed them to shift more quickly to exploit a rewarding option when they found one and only explored when the current option failed to provide valuable rewards.

Together, these results demonstrate that while the overall performance was similar, males and females tended to adopt different strategies for interacting with the same uncertain environment. The difference in explore-exploit strategies across sexes may provide us insight into potential sex-modulated mechanisms that are implicated in learning under uncertainty.

Our major finding that males learned less during exploration and explored for longer is consistent with two explanations. First, males learned more slowly during exploration and as a result, they had to explore for longer to learn which option was worth exploiting. Another possibility is that males got "stuck" in extended periods of exploration that prohibited them from applying the knowledge they have learned and committing to a rewarding option. In this view, it takes some energy to stop exploring and transition to exploit. As we have shown in our result, males had "deeper" exploratory states and they were more likely to keep exploring once started. These two explanations are not mutually exclusive because changes in learning could contribute to changes in the stickiness of exploration and vice versa. This essentially presents a chicken and egg problem. It is difficult to distinguish from behavior alone whether slower learning drives longer exploration, or vice versa, if being stuck in exploration results in slower learning. Neural measures during explore and exploit choices across sexes may help us differentiate learning signals from signals that drive exploration, and whether these signals are sex-different.

Answering these neural questions will require a way to reliably identify latent exploration and exploitation states. The Hidden Markov model (HMM) has been used to infer trial-by-trial exploration and exploitation in non-human primates (Ebitz et al., 2018). The HMM inferred latent goal states explained more variance in neural activity than other decision-variables (Ebitz et al., 2018). In our data, we have shown for the first time that the HMM model was able to label a meaningful exploratory state that matches normative definitions of exploration in the mouse model. In the future, this computational tool can complement neurobiological recording techniques to examine for neural correlates of exploration on a trial-by-trial basis, and permit the visualization of dynamic landscapes of choice behavior across individuals (as in **Figure 2C**) or with pharmacological or other challenges (Ebitz et al., 2019).

Reinforcement learning (RL) models have also been used in the past to identify levels of exploration across individuals (Daw et al., 2006; Pearson et al., 2009). However, our findings indicate that multiple latent parameters can influence how much these models explore. Here, we

found that differences in exploration between the sexes were due to differences in learning rates, not due to differences in the decision noise parameter, which is more commonly associated with exploration. While RL models are helpful for understanding cognitive or computational mechanisms, they are limited in their ability to identify when exploration is happening. The HMM model, conversely, provides no insight into mechanisms, but can tell us precisely when exploration is occurring, both in animal behaviors and RL models (Ebitz et al., 2018). By combining the HMM and RL approaches, we capitalized on the advantages of both frameworks: linking changes in exploration across sexes to underlying mechanisms of exploration. Since the broader reinforcement learning model framework is highly adaptive and amenable, in the future, the HMM and kinds of model-free analyses we completed here could also inform the design of RL models to more precisely capture explore-exploit decisions.

Future work is needed to understand the neurobiological bases of these observations. One neuromodulator that is implicated in reinforcement learning, including the transition between exploration and exploitation, *and* is strongly sex-modulated is dopamine (Beeler et al., 2010; Jenni et al., 2017). Studies have shown that dopamine signals regulate exploration via mechanisms of action selection and learning (Beeler et al., 2010; Frank et al., 2009; Humphries et al., 2012). However, due to the exclusive use of males in many foundational experiments (Beeler et al., 2010; Cinotti et al., 2019), the fact that dopamine function on decision making is strongly modulated by sex and sex-linked mechanisms is often overlooked. For example, estradiol has been demonstrated to exert both acute and chronic modulatory effects on dopamine release, reuptake, and receptor binding (Yoest et al., 2014), allowing enhanced DA release and faster DA clearance in females (Becker, 1999). This mechanism could contribute to increased reward learning observed in females during exploration. The prefrontal cortex (PFC), which receives dopamine projections, is a target brain region to understand exploration and exploitation (Ebitz et al., 2018; Jenni et al., 2017). Our previous study implicated PFC in the differences in learning strategy between males and females (Chen et al., 2020). It is possible that prefrontal cortical dopamine is particularly engaged in implementing explore-exploit strategies via sex-biased mechanisms of learning.

A frequent use of rodent operant testing is to perform cognitive assessments with similar rule structures in animals and dependent measures as are used in human tasks, to permit

translational work in animals exploring pharmacology, genetics, and other potential biological contributors to behavior (Grissom and Reyes, 2018; Heath et al., 2016). However, the vast majority of these tasks were not designed with computational models in mind, and as a result, we are unable to assess whether similar latent cognitive variables are influencing behavior in humans and rodents. Here, we establish that explore-exploit decisions are shared cognitive processes across species, making this restless bandit task an ideal tool for translational research. Parallel approaches in humans have been used to examine the explore-exploit strategic phenotype of neuropsychiatric disorders, including ADHD, addiction, and depression.(Addicott et al., 2020, 2012; Beeler et al., 2012; Blanco et al., 2013). The computational modeling used here permits fine grained quantification of individual variability in latent parameters that capture adaptive changes in exploration in changing environments. The computational approaches we develop here could help identify behavioral endophenotypes across species underlying a variety of neuropsychiatric disorders, and open up new avenues for understanding as well as rescuing dysfunction in value-based decision making.

Figure 1

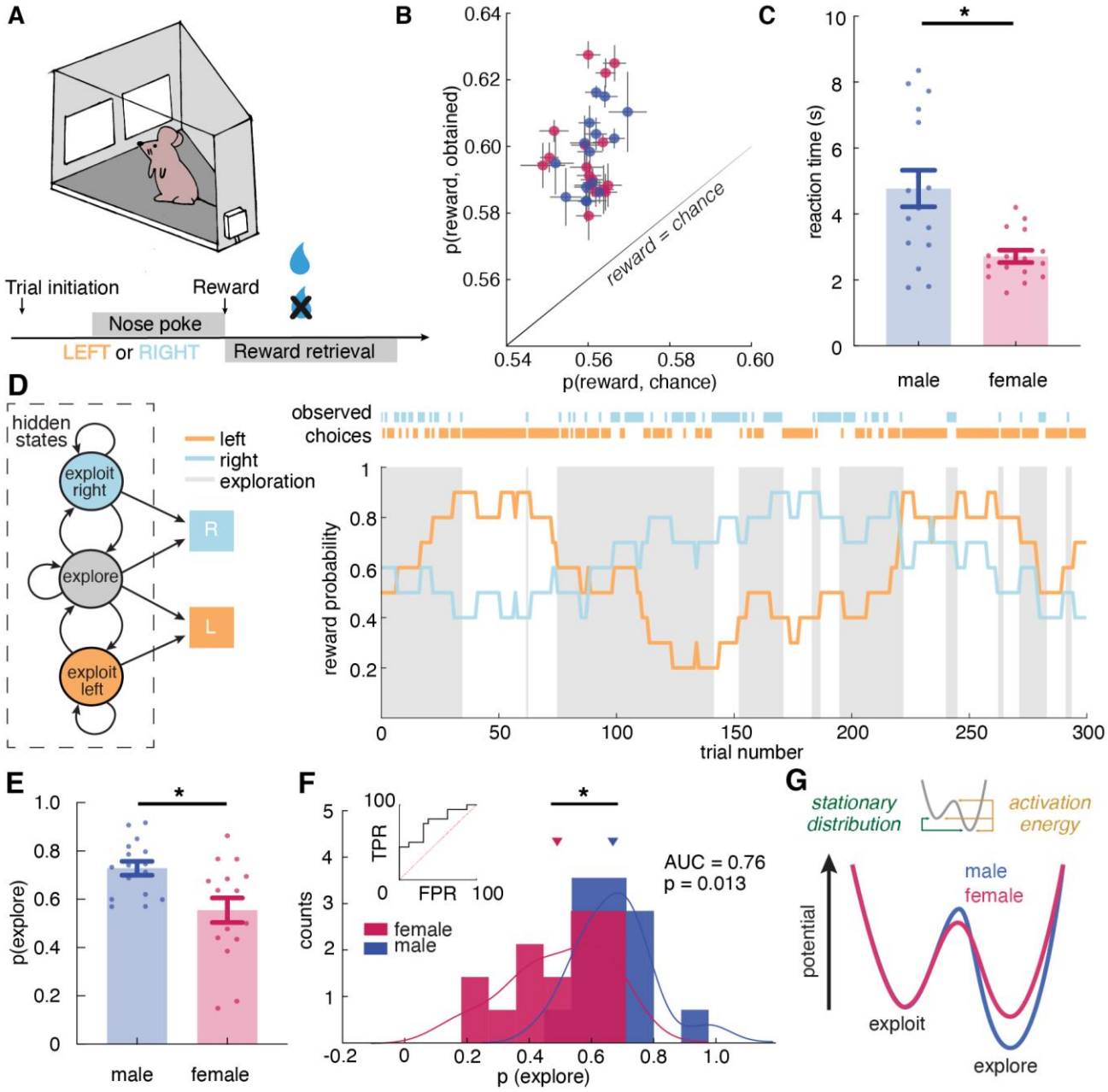


Figure 1. Male and female mice showed different exploratory strategies in a restless bandit task - males explored more than females, and they explored for longer periods of time once started.

A) Schematic of the mouse touchscreen chamber with the restless two-armed bandit task and trial structure. **B)** Average probability of obtaining reward compared to the chance

probability of reward across individuals (dots). **C)** Average reaction time of both sexes. Females responded significantly faster than did males. **D)** (left) A hidden Markov model that labeled exploration and exploitation as latent goal states underlying observed choices. This model includes an exploitation state for each arm and an exploration state where the subject chooses one of the arms randomly. (right) Reward probabilities (lines) and choices (dots) for 300 example trials for a given mouse. Shaded areas highlight explore-labeled choices. **E, F)** Average (E) and distribution (F) of percentage of Hidden Markov Model (HMM)-labeled exploratory trials in females and males. **G)** Dynamic landscape of the fitted HMMs for males and females. The model fit to males had deeper exploratory states, with higher activation energy between the states. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

Figure 2

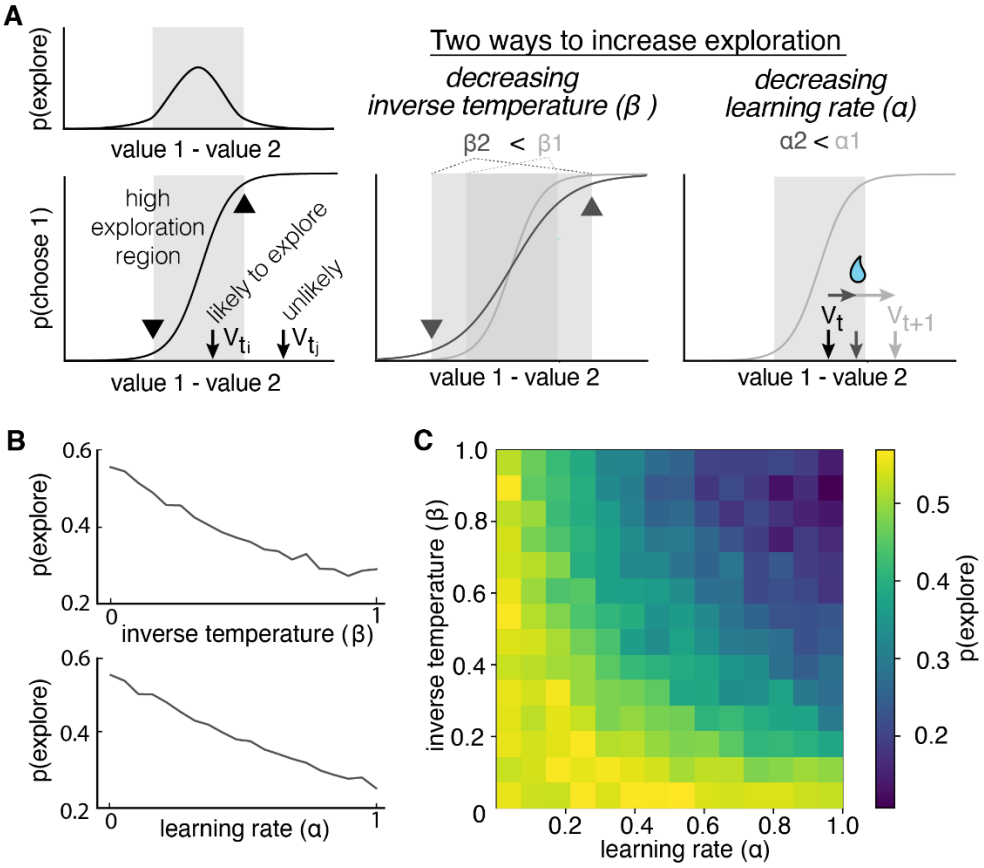


Figure 2. Multiple reinforcement learning parameters can influence the probability of exploration.

A) Exploration occurs most often when option values are close together, illustrated by the gray shaded boxes in the value-choice functions. Both decreasing inverse temperature (β) and decreasing learning rate increases exploration because each manipulation changes the amount of time spent in the high exploration zone, although the mechanisms are different. Decreasing inverse temperature (β) widens the zone by flattening the value-choice function and increasing decision noise. Decreasing learning rates (α) keeps learners in the zone for longer. **B)** Probability of exploration from 10,000 different reinforcement learning agents performing this task, initialized at different random combinations of inverse temperatures (β) and learning rates (α). Marginal relationships between decision noise (top) and learning rate (bottom) are shown here. **C)** Heatmap of all pairwise combinations of learning rate and inverse temperature.

Figure 3

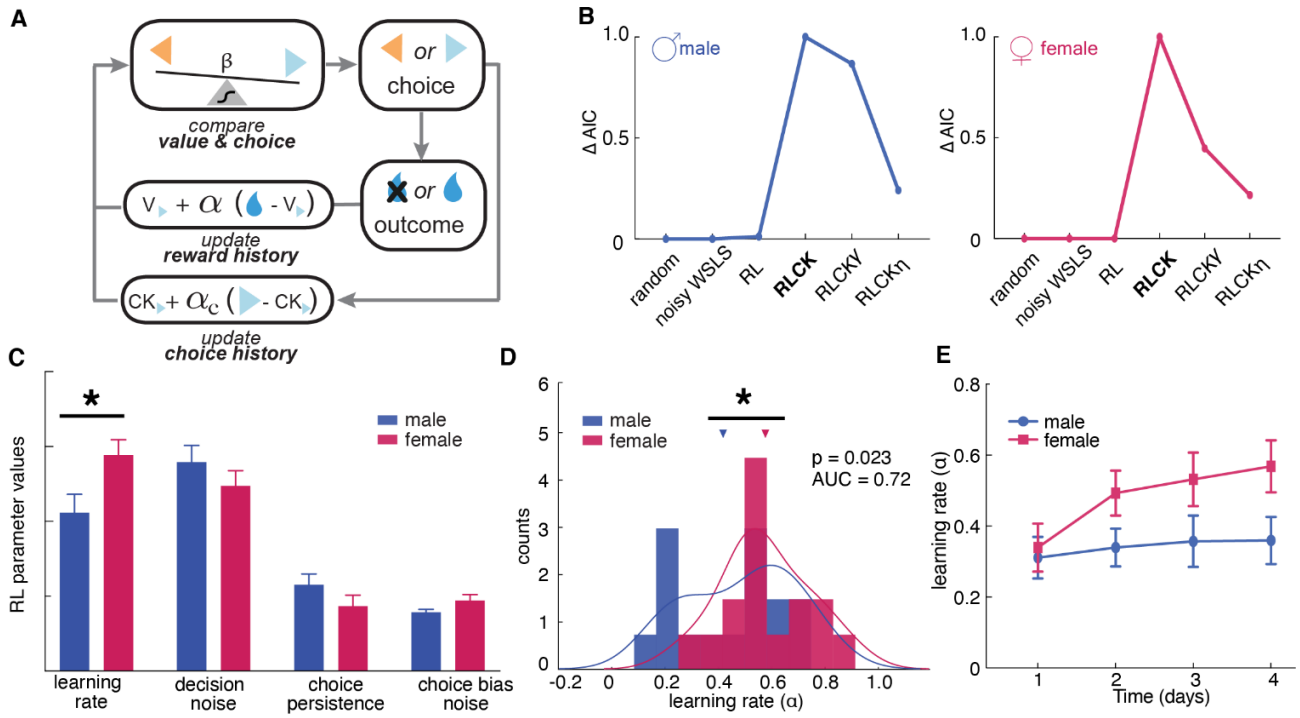


Figure 3. Sex differences in learning rate, but not decision noise, drove differences in explore-exploit decisions.

A) A diagram of latent parameters that capture learning (α), bias (α_c), inverse temperature (β) in reinforcement learning models. The models tested used a combination of these parameters (see Methods). **B)** Model comparison across 6 reinforcement learning models with various parameter combinations. The four-parameter reinforcement learning-choice kernel (RLCK) model has the highest relative likelihood in both males and females. **C)** All four parameters in the best fit RLCK model across sexes. Learning rate (α) was significantly higher in females than males. **D)** Distribution of learning rate across sexes. **E)** Learning rate in females increased across days, suggestive of meta learning. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

Figure 4

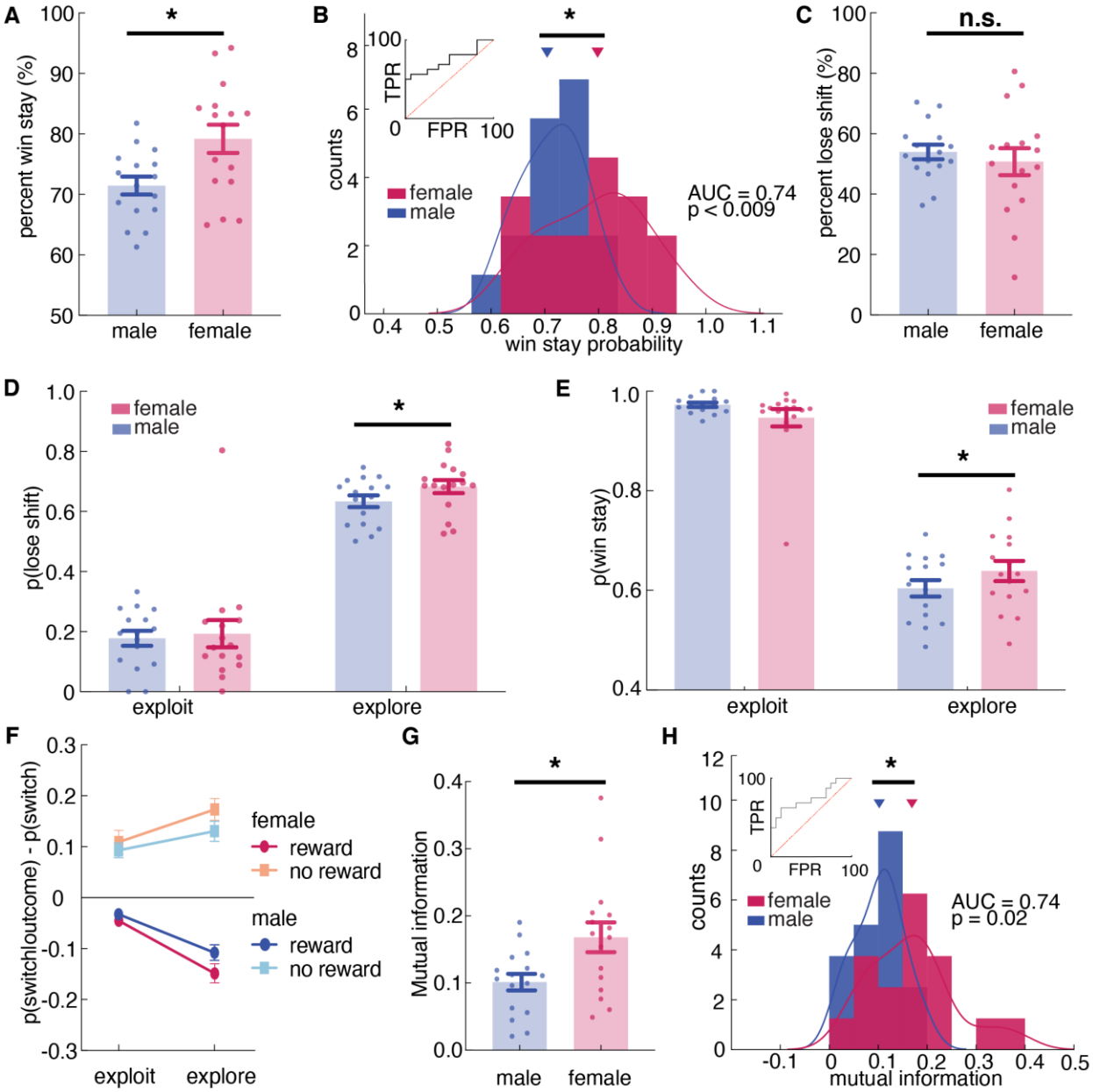
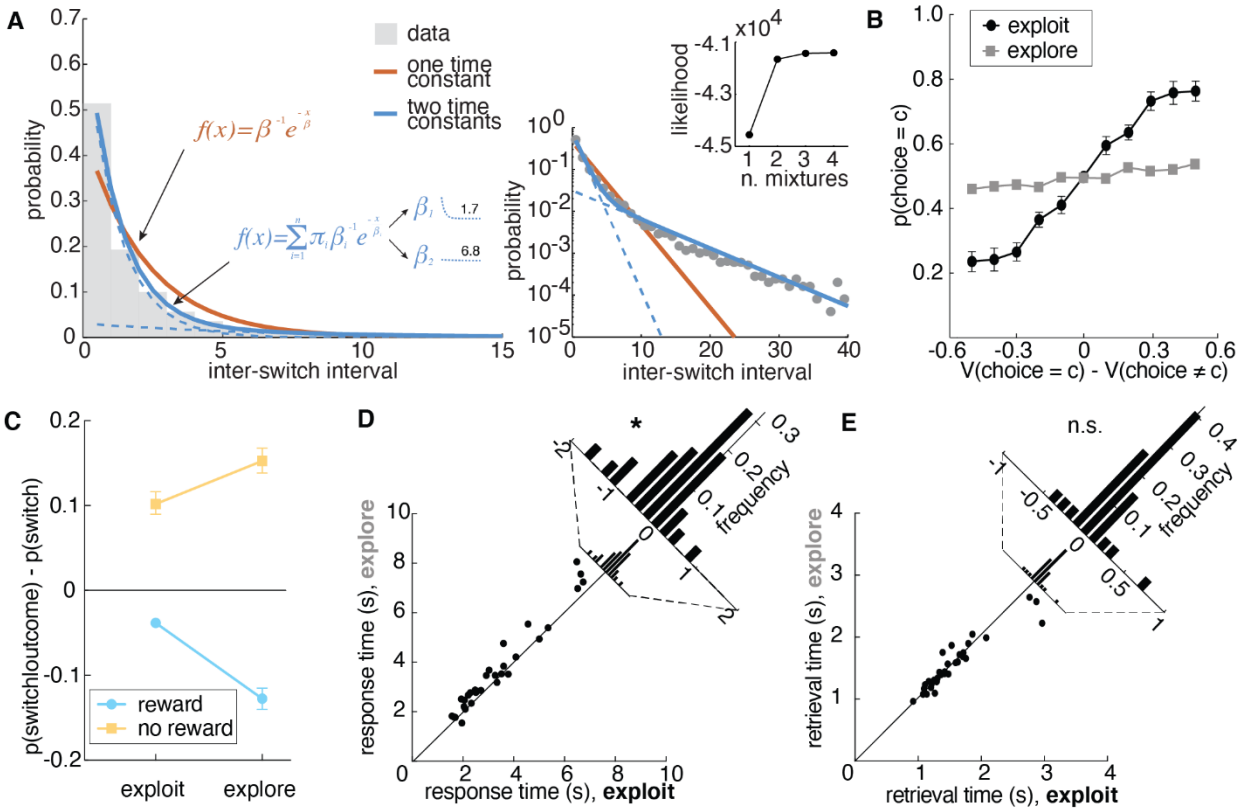


Figure 4. Females used more information about past outcomes and past choices to make decisions, and learned more during exploration.

A, B) Percent win stay behavior (A: average; B: distribution) reveals that females were more likely to stay with the same choice after a reward. **C)** Average percentage of lose shift behavior across sexes. **D)** Probability of shifting after a loss during an explore or exploit trials. **E)** Probability of staying after a win during an explore or exploit trials. **F)** during explore and exploit trials. The probability of males and females switching targets on the next trial, given the

current trial's outcome and latent state. Females only learned more during exploratory trials. **G, H)** Average (G) and distribution (H) of percentage of mutual information across all trials in females and males reveals that females use more information about past trials (choice and outcome) in making future decisions. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

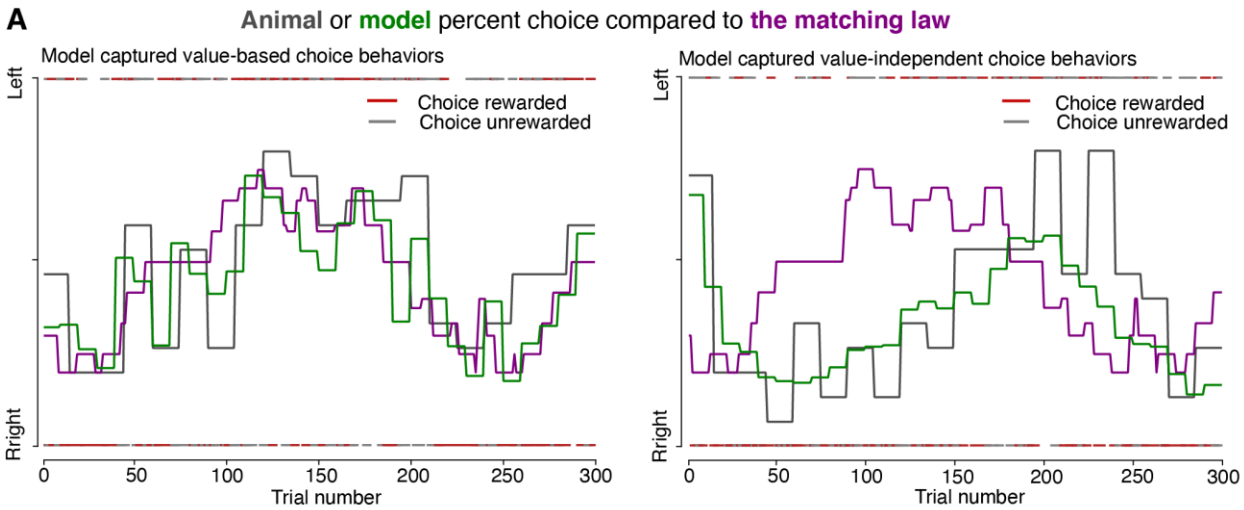
Supplemental Figures



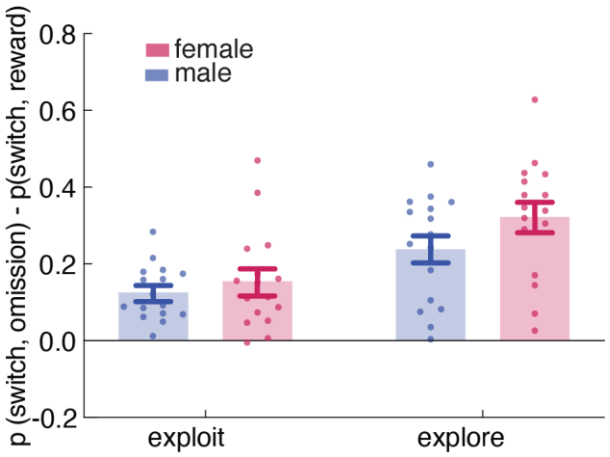
Supplemental Figure 1. Two time constants combined best describe the rate of switching choices in animals' choice behavior and Hidden Markov model validation. Related to Figure 1D.

A) The distribution of times between switch decisions (inter-switch intervals). A single probability of switching would produce exponentially distributed inter-switch intervals. Orange line, the maximum likelihood fit for a single discrete exponential distribution. Solid blue line, a mixture of two exponential distributions, with each component distribution in dotted blue. The two components reflect one fast-switching time constant (average interval, 1.7 trials) and one persistent time constant (6.8 trials). The right plot is the same as the left, but with a log scale. Inset is the log likelihood of mixtures of different numbers of exponential distributions. **B)** Probability of choice as a function of value differences between choices for exploratory and exploitative states. **C)** The probability of animals switching targets on the next trial, given the current trial's outcome and latent state. **D)** Difference in choice response time between explore and exploit choices. **E)** Difference in choice response time between explore and exploit choices. There is no significant difference in retrieval time between two latent states, suggesting that

exploration was not merely disengagement from the task. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.



Supplemental Figure 2. The best fit model, the four-parameter reinforcement learning-choice kernel (RLCK) model, captured both value-dependent and value-independent choice behaviors. Actual choices (gray) and simulated choices (green) from the best fit model (RLCK) of two example animals. Predictions from the matching law are illustrated as a contrast to the best-fitting RL model.



Supplemental Figure 3. Reward learning in explore vs. exploit state across sexes. Related to Figure 4F. The probability of males and females switching targets on the next trial, given the current trial's outcome and latent state. Females showed increased reward learning only during exploratory state.

Table 1: Main effects and interaction effects (pairwise and 3-way) of sex, outcome, latent state, and subject identity.

Table 1: Four-way Repeated Measures ANOVA

Effect	Sum Sq.	df.	Mean Sq.	F	p value
sex	0.0782	1	0.0782	6.89	0.0088
subject(sex)	2.1709	30	0.0904	7.95	< 0.00001
reward	13.1509	1	13.1509	1157.63	< 0.00001
state	35.4862	1	35.4862	3123.74	< 0.00001
sex x reward	0.2094	1	0.2094	18.44	< 0.00001
sex x state	0.0383	1	0.0383	3.37	0.0667
subject(sex) x reward	3.616	30	0.1205	10.61	< 0.00001
subject(sex) x state	1.8529	30	0.0618	5.44	< 0.00001
reward x state	0.4652	1	0.4652	40.95	< 0.00001
sex x reward x state	0.0463	1	0.0463	4.07	0.0438
subject(sex) x reward x state	0.5573	30	0.0186	1.64	0.0177
error	9.872	869	0.0114		
total	68.5207	996			

Methods

Animals. Thirty-two BL6129SF1/J mice (16 males and 16 females) were obtained from Jackson Laboratories (stock #101043). Mice arrived at the lab at 7 weeks of age, and they were housed in groups of four with ad libitum access to water while being mildly food restricted (85-95% of free feeding weight) for the experiment. Animals engaging in operant testing were housed in a 0900–2100 hours reversed light cycle to permit testing during the dark period. Before operant chamber training, animals were food restricted to 85%-90% of free feeding body weight. Operant testing occurred five days per week (Monday-Friday). All animals were cared for according to the guidelines of the National Institution of Health and the University of Minnesota.

Apparatus. Sixteen identical triangular touchscreen operant chambers (Lafayette Instrument Co., Lafayette, IN) were used for training and testing. Two walls were black acrylic plastic. The third wall housed the touchscreen and was positioned directly opposite the magazine. The magazine provided liquid reinforcer (Ensure) delivered by a peristaltic pump, typically 7ul (280 ms pump duration). ABET-II software (Lafayette Instrument Co., Lafayette, IN) was used to program operant schedules and to analyze all data from training and testing.

Behavioral task. *Two-armed spatial restless bandit task.* Animals were trained to perform a two-armed spatial restless bandit task in the touchscreen operant chamber. Each trial, animals were presented with two identical squares on the left and right side of the screen. Nose poke to one of the target locations on the touchscreen was required to register a response. Each location is associated with some probability of reward, which changes independently over time. For every trial, there is a 10% chance that the reward probability of a given arm will increase or decrease by 10%. All the walks were generated randomly with a few criteria: 1) the overall reward probabilities of two arms are within 2% of each other, preventing one arm being overly better than the other, 2) the reward probability cannot go down to 0% or go up to 100%, 3) there are no 30 consecutive trials where the reward probabilities of both arms are lower than 20% to ensure motivation. Animals ran a simple deterministic schedule on Monday to re-adapt to operant chamber after weekends off and ran a different restless bandit task each day from Tuesday to Friday. Animals ran for 2 rounds of 4 consecutive days and within each day, animals completed

either 300 trials or spent a maximum of two hours in the operant chamber. Data was recorded by the ABET II system and was exported for further analysis. All computational modeling was conducted using python.

Data analysis

General analysis techniques. Data was analyzed with custom PYTHON, MATLAB, and Prism 8 scripts. Generalized linear models, ANOVA, and t-test were used to determine sex differences over time, unless otherwise specified. P values were compared against the standard $\alpha = 0.05$ threshold. The sample size is $n = 16$ for both males and females for all statistical tests. No animal was excluded from the experiment. One outlier was removed in one analysis using ROUT method (with Q set to 1%). This outlier was from the animal that ran the lowest number of total trials. Statistics for both no outlier removal and outlier removal were reported in the result. All statistical tests used and statistical details were reported in the results. All figures depict mean ± SEM.

Mixture model. We first asked whether there were different behavioral dynamics that might correspond to exploration and exploitation. Exploration and exploitation take place on different time scales. In RL agents, for example, exploration is typically implemented via adding noise or indeterminacy to a decision-rule. The identity of choices that are caused by this noise—the exploratory choices—will thus switch more frequently than the identity of choices that depend on option value. We should see short runs of exploratory choices and long runs of exploitative ones (Ebitz et al., 2018). To the extent that choice runs end probabilistically (an assumption of the HMM framework), choice run durations (inter-switch intervals) will be exponentially distributed (Berg, 1993). To the extent that there are multiple causal regimes (such as exploration and exploitation), inter-switch intervals will be distributed as a mixture of multiple exponential distributions (**Supplemental Figure 1**). Because trials are discrete, rather than continuous, we fit mixtures of the discrete equivalent to the exponential distribution, the geometric distribution. We examined the distribution of 24,836 interswitch intervals. Adding a second mixing distribution significantly improved model fit (1-component, 1-parameter mixture log-likelihood: -44555, 2-component, 3-parameter: -41656; likelihood ratio test, $p < 10^{-32}$). Adding additional mixing distributions continued to improve model fit, a common observation in

mixture modeling. However, the continued improvement was substantially less than the leap from 1 to 2 components (**Supplemental Figure 1**; 3-component: -41431, 4-component: -41412) and additional mixtures beyond 2 had weights below 3%. This suggests that a mixture of one fast-switching regime and one slow-switching regime was the most parsimonious explanation for the data (McLachlan and Peel, 2004).

Hidden Markov model (HMM). In order to identify when animals were exploring or exploiting, we fit an HMM. This differs from the approach used in some previous studies, where reinforcement learning (RL) models were used to quantify exploration (Cinotti et al., 2019; Daw et al., 2006; Ishii et al., 2002; Jepma and Nieuwenhuis, 2011; Pearson et al., 2009). These studies typically took one of two approaches. In the first, individual variability in exploration is inferred from variability in a specific RL model parameter thought to be linked to exploration (i.e. inverse temperature/decision-noise (Cinotti et al., 2019; Ishii et al., 2002; Schweighofer and Doya, 2003). However, changing decision-noise is not the only way to change how often an RL agent explores (**Figure 2**), meaning that the inverse temperature/decision-noise parameter does not uniquely measure exploration. In the second approach, an RL model is fit to behavior and decisions that deviate from model values are labeled as exploratory (Daw et al., 2006; Jepma and Nieuwenhuis, 2011; Pearson et al., 2009). This approach is based on the rationale that exploration is a non-reward maximizing goal. However, a non-reward maximizing goal would produce choices that are *orthogonal* to reward, not the errors of reward maximization identified by this method (Averbeck et al., 2017; Ebitz et al., 2018). Beyond issues of face validity, both approaches are sensitive to any misspecifications in the behavioral model: a subject who deviates more from the assumptions of the model may seem to explore more since the subject will also have high decision-noise and a large number of choices will not match the model's predictions. For these reasons, we chose to apply an HMM approach, which models exploration directly, rather than inferring it from collateral observations, and does not depend on any kind of generative model of how the animals perform the task--a salient concern here, where we suspect that these computations differ between the sexes.

In an HMM framework, choices (y) are "emissions" that are generated by an unobserved decision process that is in some latent, hidden state (z). Latent states are defined by both the probability of making each choice (k, out of N_k possible options), and by the probability of

transitioning from each state to every other state. Our model consisted of two types of states, the explore states and the exploit state. The emissions model for the explore state was uniform across the options:

$$p(y_t = k | z_t = explore) = \frac{1}{N_k}$$

This is simply the maximum entropy distribution for a categorical variable and does not imply or enforce that exploratory choices were random choices (Ebitz et al., 2020, 2019). Because exploitation involves repeated samples to each option, exploit states only permitted choice emissions that matched one option. That is:

$$\begin{cases} p(y_t = k | z_t = exploit_i, k \in exploit_i) = 1 \\ p(y_t = k | z_t = exploit_i, k \notin exploit_i) = 0 \end{cases}$$

The latent states in this model are Markovian meaning that they are time-independent. They depend only on the most recent state (z_t):

$$p(z_t | z_{t-1}, y_{t-1}, \dots, z_1, y_1) = p(z_t | z_{t-1})$$

This means that we can describe the entire pattern of dynamics in terms of a single transition matrix. This matrix is a system of stochastic equations describing the 1-time-step probability of transitioning between every combination of past and future states (i, j).

$$p(z_t = i | z_{t-1} = j)$$

Here, there were 3 possible states (2 exploit states and 1 explore state) but parameters were tied across exploit states such that each exploit state had the same probability of beginning (from exploring) and of sustaining itself. Transitions out of the exploration, into exploitative states, were similarly tied. The model also assumed that the mice had to pass through exploration in order to start exploiting a new option, even if only for a single trial. Through fixing the emissions model, constraining the structure of the transmission matrix, and tying the parameters, the final HMM had only two free parameters: one corresponding to the probability of exploring, given exploration on the last trial, and one corresponding to the probability of exploiting, given exploitation on the last trial.

The model was fit via expectation-maximization using the Baum Welch algorithm (Bilmes, 1998). This algorithm finds a (possibly local) maxima of the complete-data likelihood. A complete set of parameters θ includes the emission and transition models, discussed already,

but also initial distribution over states, typically denoted as PI. Because the mice had no knowledge of the environment at the first trial of the session, we assumed they began by exploring, rather than add another parameter to the model here. The algorithm was reinitialized with random seeds 20 times, and the model that maximized the observed (incomplete) data log likelihood across all the sessions for each animal was ultimately taken as the best. To decode latent states from choices, we used the Viterbi algorithm to discover the most probable a posteriori sequence of latent states.

Analyzing model dynamics. In order to understand how exploration and exploitation changed across males and females, we analyzed the HMMs. The term "dynamics" means the rules or laws that govern how a system evolves over time. Here, the system of interest was decision making behavior, specifically at the level of the hidden explore and exploit goals. In fitting our HMMs, we were fitting a set of equations that describe the dynamics of these goals: the probability of transitions between exploration and exploitation and vice versa. Of course, having a set of fitted equations is a far cry from understanding them. To develop an intuition for how sex altered the dynamics of exploration, we therefore turned to analytical tools that allowed us to directly characterize the energetic landscape of behavior (**Figure 2C**).

In statistical mechanics, processes within a system (like a decision-maker at some moment in time) occupy states (like exploration or exploitation). States have energy associated with them, related to the long-time scale probability of observing a process in those states. A low energy state is one that is very stable and deep, much like a valley between two high peaks. Low energy states will be over-represented in the system. A high energy state, like a ledge on the side of a mountain, is considerably less stable. High energy states will be under-represented in the long-term behavior of the system. The probability of observing a process in a given state i will is related to the energy of that state (E_i) via the Boltzman distribution:

$$p_i = \frac{1}{Z} e^{\frac{-E_i}{k_B T}}$$

Where Z is the partition function of the system, k_B is the Boltzman constant, and T is the temperature of the system. If we focus on the ratio between two state probabilities, the partition functions cancel out and the relative occupancy of the two states is now a simple function of the difference in energy between them:

$$\frac{p_i}{p_j} = e^{\frac{-(E_i - E_j)}{k_B t}}$$

Rearranging, we can now express the difference in energy between two states as a function of the difference in the long-term probability of those states being occupied:

$$\ln\left(\frac{p_i}{p_j}\right) k_B T = E_j - E_i$$

Meaning that the difference in the energetic depth of the states is proportional to the natural log of the probability of each state, up to some multiplicative factor $k_B T$. To calculate the probability of exploration and exploitation (p_i and p_j), we calculated the stationary distribution of the fitted HMMs. The stationary distribution is the equilibrium probability distribution over states. This means that this distribution is the relative frequency of each state that we would observe if the model's dynamics were run for an infinite period of time. Each entry of the model's transition matrix reflects the probability that the mice would move from one state (e.g. exploring) to another (e.g. exploiting one of the options) at each moment in time. Because the parameters for all the exploitation states were tied, each transition matrix effectively had two states—an explore state and a generic exploit that described the dynamics of all exploit states. Each of the k animals had its own transition matrix (A_k), which describes how the entire system—an entire probability distribution over states—would evolve from time point to time point. We observe how the dynamics evolve any probability distribution over states (π) by applying the dynamics to this distribution:

$$\pi_{t+1} = \pi_t A_k$$

Over many time steps, ergodic systems will reach a point where the state distributions are unchanged by continued application of the transition matrix as the distribution of states reaches its equilibrium. That is, in stationary systems, there exists a stationary distribution, π^*, such that:

$$\pi^* = \pi^* A_k$$

If it exists, this distribution is a (normalized) left eigenvector of the transition matrix A_k with an eigenvalue of 1, so we solved for this eigenvector to determine the stationary distribution of each A_k. Whe then took an average of these stationary distributions within each sex, plugged them back into the Boltzman equations to calculate the relative energy (depth) of exploration and exploitation as illustrated in **Figure 2C**.

In order to understand the dynamics of our coarse-grained system, we need to not only understand the depth of the two states, but also the height of the energetic barrier between them: the activation energy required to transition from exploration to exploitation and back again. Here, we build on an approach from chemical kinetics that relates the rate of transition between different conformational states to the energy required to affect these transitions. The Arrhenius equation relates the rate of transitions away from a state (k) to the activation energy required to escape that state (E_a):

$$k = Ae^{\frac{E_a}{k_B T}}$$

Where A is a constant pre-exponential factor related to the readiness of reactants to undergo the transformation. We will set this to 1 for convenience. Again, $k_B T$ is the product of temperature and the Boltzman constant. Note the similarities between this equation and the Boltzman distribution illustrated earlier. Rearranging this equation to solve for activation energy yields:

$$E_a = -\ln\left(\frac{k}{A}\right)k_B T$$

Thus, much like the relative depth of each state, activation energy is also proportional to some measurable function of behavior, up to some multiplicative factor $k_B T$. Note that our approach has only identified the energy of three discrete states (an explore state, an exploit state, and the peak of the barrier between them). These are illustrated by tracing a continuous potential through these three points only to provide a physical intuition for these effects.

<u>Reinforcement learning models.</u> We fitted six reinforcement learning (RL) models that could potentially characterize animals' choice behaviors, with details of each RL model as below. To identify the model that best captured the computations used by the animals, we compared model fits across six reinforcement learning models with different combinations of latent parameters. AIC weights were calculated from AIC values of each model for each sex and compared across models to determine the best model with the highest relative likelihood.

The first model assumes that animals choose between two arms randomly with some overall bias for one side over the other. This choice bias for choosing left side over right side is captured with a parameter *b*. The probability of choosing left side on trial *t* is:

$$p_t^L = b \qquad [1] \quad \textit{"random"}$$

The second model is a noisy win-stay lose-shift (WSLS) model that adapts choices with regards to outcomes. This model assumes a win-stay lose-shift policy that is to repeat a rewarded choice and to switch to the other choice if not rewarded. Furthermore, this model includes a parameter ϵ that captures the level of randomness, allowing a stochastic application of the win-stay lose-shift policy. The probability of choosing arm *k* on trial *t* is:

$$p_t^k = \begin{cases} 1 - \frac{\epsilon}{2}, if\ (c_{t-1} = k\ and\ r_{t-1} = 1\ OR\ c_{t-1} \neq k\ and\ r_{t-1} = 0) \\ \frac{\epsilon}{2}, if\ (c_{t-1} \neq k\ and\ r_{t-1} = 1\ OR\ c_{t-1} = k\ and\ r_{t-1} = 0) \end{cases} \quad [2]\ \textit{“noisy WSLS”}$$

c_t indicates the choice on trial *t* and r_t is a binary variable that indicates whether or not trial *t* was rewarded.

The third model is a basic delta-rule reinforcement learning (RL) model. This two-parameter model assumes that animals learn by consistently updating Q values, which are values defined for options (left and right side). These Q values, in turn, dictate what choice to make next. For example, in a multi-armed bandit task, Q_t^k is the value estimation of how good arm *k* at trial *t*, and is updated based on the reward outcome of each trial:

$$Q_{t+1}^k = Q_t^k + \alpha(r_t - Q_t^k) \qquad [3]\ \textit{”RL”}$$

In each trial, $r_t - Q_t^k$ captures the reward prediction error (RPE), which is the diffe-rence between expected outcome and the actual outcome. The parameter *a* is the learning rate, which determines the rate of updating RPE. With Q values defined for each arm, choice selection on each trial was performed based on a Softmax probability distribution:

$$p(a_{t+1} = k) = \frac{e^{\beta Q_t^k}}{\sum_j e^{\beta Q_t^j}}$$

, where inverse temperature β determines the level of random decision noise.

The fourth model incorporates a choice updating rules in addition to the value updating rule in model 3. The model assumes that choice kernel, which captures the outcome-independent tendency to repeat a previous choice, also influences decision making. The choice kernel updating rule is similar to the value-updating rule:

$$CK_{t+1}^k = CK_t^k + \alpha_c(a_t^k - CK_t^k) \qquad [4]\ \textit{“RLCK”}$$

, where a_t^k is a binary variable that indicates whether or not arm *k* was chosen on trial *t* and a_t is choice kernel updating rate, characterizing choice persistence. The value and choice kernel term were combined to compute the probability of choosing arm *k* on trial *t*:

$$p_t^k = \frac{e^{(\beta Q_t^k + \beta_c CK_t^k)}}{\sum_j e^{(\beta Q_t^j + \beta_c CK_t^j)}}$$

, where β_c is the inverse temperature associated with the choice kernel, capturing stickiness of choice.

The fifth model is the same with the fourth model, except that this model includes another parameter γ that modulates learning rate when the choice is not rewarded. This model assumes that the learning rate is different for rewarded and unrewarded trials.

$$Q_{t+1}^k = \begin{cases} Q_t^k + \alpha\left(r_t - Q_t^k\right), r_t = 1 \\ Q_t^k + \gamma \times \alpha\left(r_t - Q_t^k\right), r_t = 0 \end{cases} \qquad [5]\ \textit{“RLCK}\gamma\textit{”}$$

The sixth model is also similar to model 4, except that this model includes another parameter η that tunes the balance between the value updating rule and the choice kernel updating rule. This model assumes that animals could be using two policies (value and choice kernel) to different extent, that is some animals could depend their choices more heavily on values and some animals could be more dependent on choice preference. In this model, the probability of choosing arm *k* on trial *t*:

$$p_t^k = \frac{e^{(\eta\beta Q_t^k + (1-\eta)\beta_c CK_t^k)}}{\sum_j e^{(\eta\beta Q_t^j + (1-\eta)\beta_c CK_t^j)}} \qquad [6]\ \textit{“RLCK}\eta\textit{”}$$

<u>Conditional mutual information.</u> We quantified the extent to which choice history was informative about current choices as the conditional mutual information between the current choice (C) and the last choice (C_t-1), conditioned on the reward outcome of the last trial (R):

$$I(C_t; C_{t-1}|R) = \sum_{r \in R} \sum_{c_{t-1} \in C} \sum_{C_t \in C} p_{C_t, C_{t-1}, R}(c_t, c_{t-1}, r) log \frac{p_R(r) p_{C_t, C_{t-1}, R}(c_t, c_{t-1}, r)}{p_{C_t, R}(c_t, r) p_{C_{t-1}, R}(c_{t-1}, r)}$$

, where the set of choice options (C) represented the 2 options (left/right).

Competing Interests

The authors have declared no competing interests.

Acknowledgements

This work was supported by NIMH R01 MH123661, NIMH P50 MH119569, and NIMH T32 training grant MH115886, startup funds from the University of Minnesota (N.M.G.), a Young Investigator Grant from the Brain and Behavior Foundation (R.B.E.), an Unfettered Research Grant from the Mistletoe Foundation (R.B.E.), and Fonds de Recherche du Québec Santé, Chercheur-Boursier Junior 1, #284309 (R.B.E). Thank you to Briana E. Mork and Lisa S. Curry-Pochy for help improve this manuscript.

Authors Contributions

Conceptualization, N.M.G., R.B.E., and C.S.C.; Methodology, C.S.C., R.B.E., and N.M.G.; Investigation, C.S.C. E.K. and A.H.; Writing - Original Draft, C.S.C., R.B.E. and E.K.; Writing-Review & Editing, C.S.C., R.B.E., E.K., and N.M.G.; Funding Acquisition, R.B.E. and N.M.G.; Supervision, R.B.E. and N.M.G.

References

Addicott MA, Baranger DAA, Kozink RV, Smoski MJ, Dichter GS, McClernon FJ. 2012. Smoking withdrawal is associated with increases in brain activation during decision making and reward anticipation: a preliminary study. *Psychopharmacology* **219**:563–573. doi:10.1007/s00213-011-2404-3

Addicott MA, Pearson JM, Schechter JC, Sapyta JJ, Weiss MD, Kollins SH. 2020. Attention-deficit/hyperactivity disorder and the explore/exploit trade-off. *Neuropsychopharmacology*. doi:10.1038/s41386-020-00881-8

Addicott MA, Pearson JM, Sweitzer MM, Barack DL, Platt ML. 2017. A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. *Neuropsychopharmacology* **42**:1931–1939. doi:10.1038/npp.2017.108

Averbeck B, Costa V, Eisenberg D, Czarapata J, Berman K, Murray E. 2017. 376. Subcortical Contributions to the Explore-Exploit Tradeoff. *Biological Psychiatry*. doi:10.1016/j.biopsych.2017.02.393

Becker JB. 1999. Gender differences in dopaminergic function in striatum and nucleus accumbens. *Pharmacol Biochem Behav* **64**:803–812. doi:10.1016/s0091-3057(99)00168-9

Beeler JA, Daw N, Frazier CRM, Zhuang X. 2010. Tonic dopamine modulates exploitation of reward learning. *Front Behav Neurosci* **4**:170. doi:10.3389/fnbeh.2010.00170

Beeler JA, Frazier CRM, Zhuang X. 2012. Putting desire on a budget: dopamine and energy expenditure, reconciling reward and resources. *Front Integr Neurosci* **6**:49. doi:10.3389/fnint.2012.00049

Berg HC. 1993. Random Walks in Biology. Princeton University Press.

Bilmes J. 1998. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.

Blanco NJ, Otto AR, Maddox WT, Beevers CG, Love BC. 2013. The influence of depression symptoms on exploratory decision-making. *Cognition* **129**:563–568. doi:10.1016/j.cognition.2013.08.018

Chen CS, Becket Ebitz R, Bindas SR, David Redish A, Hayden BY, Grissom NM. 2020. Divergent Strategies for Learning in Males and Females. *Curr Biol* **0**. doi:10.1016/j.cub.2020.09.075

Cinotti F, Fresno V, Aklil N, Coutureau E, Girard B, Marchand AR, Khamassi M. 2019. Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Sci Rep* **9**:6770. doi:10.1038/s41598-019-43245-z

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. 2006. Cortical substrates for exploratory decisions in humans. *Nature* **441**:876–879. doi:10.1038/nature04766

Doya K. 2002. Metalearning and neuromodulation. *Neural Netw* **15**:495–506. doi:10.1016/s0893-6080(02)00044-8

Ebitz RB, Albarran E, Moore T. 2018. Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. *Neuron* **97**:475. doi:10.1016/j.neuron.2018.01.011

Ebitz RB, Sleezer BJ, Jedema HP, Bradberry CW, Hayden BY. 2019. Tonic exploration governs both flexibility and lapses. *PLoS Comput Biol* **15**:e1007475. doi:10.1371/journal.pcbi.1007475

Ebitz RB, Tu JC, Hayden BY. 2020. Rules warp feature encoding in decision-making circuits. *PLoS Biol* **18**:e3000951. doi:10.1371/journal.pbio.3000951

Filipowicz ALS, Levine J, Piasini E, Tavoni G, Kable JW, Gold JI. 2019. The complexity of model-free and model-based learning strategies. doi:10.1101/2019.12.28.879965

Frank MJ, Doll BB, Oas-Terpstra J, 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

Green T, Flash S, Reiss AL. 2019. Sex differences in psychiatric disorders: what we can learn from sex chromosome aneuploidies. *Neuropsychopharmacology* **44**:9–21. doi:10.1038/s41386-018-0153-2

Grissom NM, Reyes TM. 2018. Let's call the whole thing off: evaluating gender and sex differences in executive function. *Neuropsychopharmacology*. doi:10.1038/s41386-018-0179-5

Heath CJ, Phillips BU, Bussey TJ, Saksida LM. 2016. Measuring Motivation and Reward-Related Decision Making in the Rodent Operant Touchscreen System. *Curr Protoc Neurosci* **74**:8.34.1–20. doi:10.1002/0471142301.ns0834s74

Humphries MD, Khamassi M, Gurney K. 2012. Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. *Front Neurosci* **6**:9. doi:10.3389/fnins.2012.00009

Ishii S, Yoshida W, Yoshimoto J. 2002. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Netw* **15**:665–687. doi:10.1016/s0893-6080(02)00056-4

Jenni NL, Larkin JD, Floresco SB. 2017. Prefrontal Dopamine D1 and D2 Receptors Regulate Dissociable Aspects of Decision Making via Distinct Ventral Striatal and Amygdalar Circuits. *J Neurosci* **37**:6200–6213. doi:10.1523/JNEUROSCI.0030-17.2017

Jepma M, Nieuwenhuis S. 2011. Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory. *J Cogn Neurosci* **23**:1587–1596. doi:10.1162/jocn.2010.21548

Kool W, McGuire JT, Rosen ZB, Botvinick MM. 2010. Decision making and the avoidance of cognitive demand. *J Exp Psychol Gen* **139**:665–682. doi:10.1037/a0020198

Kurdi B, Gershman SJ, Banaji MR. 2019. Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proc Natl Acad Sci U S A* **116**:6035–6044. doi:10.1073/pnas.1820238116

Leao D Jr, Fragoso M, Ruffino P. 2004. Regular conditional probability, disintegration of probability and Radon spaces. *Proyecciones* **23**:15–29.

Maney DL. 2016. Perils and pitfalls of reporting sex differences. *Philos Trans R Soc Lond B Biol Sci* **371**:20150119. doi:10.1098/rstb.2015.0119

McLachlan GJ, Peel D. 2004. Finite Mixture Models. John Wiley & Sons.

Pearson JM, Hayden BY, Raghavachari S, Platt ML. 2009. Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Curr Biol* **19**:1532–1537. doi:10.1016/j.cub.2009.07.048

Pisupati S, Chartarifsky-Lynn L, Khanal A, Churchland AK. 2019. Lapses in perceptual judgments reflect exploration. *Cold Spring Harbor Laboratory*. doi:10.1101/613828

Poling A, Edwards TL, Weeden M, Foster TM. 2011. The Matching Law. *Psychol Rec* **61**:313–322. doi:10.1007/BF03395762

Schweighofer N, Doya K. 2003. Meta-learning in reinforcement learning. *Neural Netw* **16**:5–9. doi:10.1016/s0893-6080(02)00228-9

Shansky RM. 2019. Are hormones a "female problem" for animal research? *Science* **364**:825–826. doi:10.1126/science.aaw7570

van den Bos R, Homberg J, de Visser L. 2013. A critical review of sex differences in decision-making tasks: Focus on the Iowa Gambling Task. *Behavioural Brain Research*.

doi:10.1016/j.bbr.2012.10.002

Wilson RC, Bonawitz E, Costa VD, Ebitz RB. 2021. Balancing exploration and exploitation with information and randomization. *Curr Opin Behav Sci* **38**:49–56. doi:10.1016/j.cobeha.2020.10.001

Wyner AD. 1978. A definition of conditional mutual information for arbitrary ensembles. *Information and Control* **38**:51–59. doi:10.1016/S0019-9958(78)90026-8

Yoest KE, Cummings JA, Becker JB. 2014. Estradiol, dopamine and motivation. *Cent Nerv Syst Agents Med Chem* **14**:83–89. doi:10.2174/1871524914666141226103135