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, Université de Montréal, 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

Abstract

Sex differences in cognitive processes could set the stage for sex-modulated 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 across sexes often show minimal differences. Computational tools allow us to uncover latent variables in reinforcement learning that define different decision making approaches, even in animals with similar correct performance. Here, we quantify sex differences in latent variables underlying behavior in a classic value-based decision-making task: a restless 2-armed bandit. While males and females had similar accuracy, they achieved this performance via different patterns of exploration. Males made more exploratory choices overall, largely because they appeared to get stuck in exploration once they had started. Females explored less, but learned more quickly when they did so. Together, these results suggest that sex exerts stronger influences on learning and decision making during periods of self-initiated exploration than during stable choices. These findings pinpoint the neural mechanisms of exploration as potentially conferring sex-biased vulnerability to addictions, neurodevelopmental disabilities, and other neuropsychiatric disorders.

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 [1–3]. 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 variable, even more so than would be expected given the non-dichotomous, overlapping impacts of sex mechanisms on cognition [4]. 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 "percent correct". This means that equivalent levels of correct performance could mask real 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 males and females can take very different strategic paths to the same outcome [5]. Here, we apply this new approach--using 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 [6,7], many of which are also sex-biased [1–3]. This implies that sex differences in exploration and exploitation could contribute to sex-linked vulnerability to

these conditions. Further, because exploration is a major source of errors in task performance more broadly [8,9], sex-differences in exploration may be accessed in complex ways by tasks and could contribute to a variety of other sex differences seen across tasks and species etc [1,10]. However, it is not yet clear whether or not there are sex differences in exploration and exploitation.

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 [11]. 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 the touch-screen operant chambers (**Figure 1A**). Animals were presented with two physically identical targets (squares) on the left and right of the screen and indicated their choices by nose poking at one of two target locations. Each location offered some probability of reward, which changed randomly and independently across trials. The dynamic reward contingencies encouraged both exploration and exploitation.

80 The animals must exploit a good option when it is found, but also occasionally explore the other
81 option, which could become better at any time. Mice performed 2 repetitions of 4 consecutive
82 sessions of the restless bandit task.

83 To examine task performance, we calculated the average probability of reward obtained
84 in males and females. Because rewards schedules were stochastic, sessions could differed
85 slightly in the amount of reward that was available. We therefore compared performance against
86 the average probability of reward if choosing randomly. Regardless of sex, mice were able to
87 earch reward more frequently than chance (**Figure 1B,** two-way ANOVA, $F(1, 60) = 228.9$,
88 $p<0.0001$). There was no significant sex difference in the probability of rewards acquired above
89 chance (**Figure 1B**, main effect of sex, $F(1, 30) = 0.05$, $p = 0.83$).

90 Similar levels of accuracy do not require or imply a similar approach to the task. Our
91 previous study suggested that males and females could achieve similar learning performance via
92 divergent learning strategies [5]. However, different strategies might take different amounts of time
93 to execute [5,12–14]. Therefore, if males and females had adopted different approaches to this task,
94 then we might expect the response time to differ between males and females, despite the
95 similarities in learning performance. Indeed, females responded significantly faster than did
96 males (**Figure 1C**, main effect of sex, $t(30) = 3.52$, $p = 0.0014$), suggesting that the computations
97 used in each sex may differ and, if so, that the females' strategy was faster than the males'.

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

100 Despite similar performance, different response times could suggest that males and
101 females employed different strategies in this task. One possibility is that the difference in choice

102 reaction time was due to sex differences in the level of exploration. Prior research has shown that
103 exploratory choices take longer than exploitative choices [8,11]. Therefore, perhaps males took
104 longer to make a choice because a greater proportion of their choices were exploratory. To test
105 this hypothesis, we used a hidden Markov model (HMM) to determine whether choices were
106 exploratory or exploitative [11].

107 The HMM allows us to make statistical inferences about the probability that each choice
108 was due to exploration or exploitation via modeling these as the latent goal states underlying
109 choice (see Methods). This approach differs from previous studies, which used reinforcement
110 learning (RL) models to quantify exploration [15–19]. Past studies have typically taken one of two
111 approaches. In the first, individual variability in exploration is inferred from variability in a
112 specific RL model parameter thought to be linked to exploration (i.e. inverse
113 temperature/decision-noise; [18–20]. As we will see shortly, this approach is flawed because
114 changing decision-noise is not the only way to change how often an RL agent explores. In the
115 second approach, an RL model is fit to behavior and decisions that deviate from model values
116 are labeled as exploratory [15–17]. This approach is based on the rationale that exploration is a non-
117 reward maximizing goal, but a non-reward maximizing goal would produce choices that are
118 *orthogonal* to reward, not errors of reward maximization [11,21]. Beyond issues of face validity,
119 both of these approaches are sensitive to any mispecifications in the behavioral model: a subject
120 who deviates more from the assumptions of the model may seem to explore more since the
121 subject will also have high decision-noise and a large number of choices will not match the
122 model's predictions. The HMM approach, conversely, models exploration directly, rather than
123 inferring it from collateral observations, and does not depend on any kind of generative model of

124 how the animals perform the task--a salient concern here, where we suspect that these
125 computations differ between the sexes.

126 The HMM method has never before been used to quantify exploration in mice, so we first
127 asked whether it was appropriate here. The HMM method works because *sequences* of
128 exploratory decisions look very different from exploitative ones, at least in reinforcement
129 learning agents and primates [11]. When agents exploit, they repeatedly sample the same option,
130 switching only very rarely. However, because exploration requires investigatory samples, runs of
131 exploratory choices tend not to repeat the same option. They tend to switch far more frequently,
132 closer to what we would expect from random samples from the environment (Ebitz, et al., 2018).
133 Therefore, if mice were alternating between exploration and exploitation in this task, we would
134 expect to see evidence of two distinct patterns of switching in their behavior. Indeed, choice run
135 durations (i.e. the distribution of inter-switch intervals) were parsimoniously described as a
136 mixture of two different patterns: one regime where choices switched quickly (mean switching
137 time = 1.7 trials, compared to random choices at 2 trials; 80% of choice runs) and one regime
138 where they changed slowly (mean switching time = 6.8 trials; **Supplemental Figure 1**). Thus,
139 mice had evidence of fast-switching (putatively exploratory) and slow-switching (putatively
140 exploitative) regimes in their behavior. Note that explore-labeled choices are more likely to also
141 be switch choices, but not all explore-labeled choices are switches, and not all exploit-labeled
142 choices are stay decisions.

143 Next, we used a simple 2-parameter HMM to infer the likelihood that each choice was
144 part of each regime: the putatively exploratory regime, or the putatively exploitative one (see
145 Methods; **Figure 1D**). To evaluate the face validity of the HMM approach, we next asked
146 whether HMM-labeled exploratory or exploitative choices matched the normative definitions of

these terms. 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 orthogonal to reward value (**Figure 1E)**. Exploit-labeled choices were a function of value difference between two choices - that is the probability of choosing the higher value choice increases as a function of the value. However, in explore-labeled choices, the average value of choices chosen during exploration was not different from chance (one sample t-test, $t(10) = 0.16$, $p = 0.87$), which suggests that HMM-labeled exploration was not driven by some goal other than reward maximization.

Second, reward learning elevates during exploration because the goal of exploration is learning about choice-outcome associations. 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 (**Figure 1F**, two-way RM ANOVA, interaction term, $F(1,31) = 51.2$, $p < 0.001$). Finally, we asked whether the labels inferred by the HMM were meaningful - that is whether two states 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 (**Figure 1G**, t test, $t(31) = 3.66$, $p = 0.0009$), which is consistent with previous findings that exploration slows down decision making [11]. 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 2**, 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 during exploration. 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 2A**). Females explored much less with only 55.4% ± 20.4% STD of trials being exploratory (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 2B,** 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 model parameter (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 2C**, 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, struggling to 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. This suggests that there were sex differences in the animals' approach to this task. However, a crucial question remained unanswered: why were males more exploratory? How did the latent computations used by males differ from those used by females? To address this question, we turned to reinforcement learning models to look for individual variability in parameters that could influence exploration and exploitation [15–18].

One feature about RL models is that multiple latent parameters can influence the probability of exploration. Exploration is a function of both the differences in values and decision noise. To illustrate, we ran simulations focused on a simple, 2-parameter RL model, with one learning rate parameter (α) and one parameter for decision noise (inverse temperature β). Commonly, only the latter parameter is thought to reflect the level of random exploration, so most previous studies of exploration have focused on this inverse temperature parameter to account for decision noise [19,22].

As illustrated in **Figure 3A**, changing the inverse temperature of the softmax function influences the probability of exploration - the larger β is, 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 the model (**Figure 3B**; GLM, main effect of inverse temperature, $\beta 1 = -0.11$, $p < 0.0001$). Another 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 3A** (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 3B**; 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 3C**; 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 and increased. Together, these results suggest that sex-differences in exploration (and perhaps individual differences in exploration more broadly) could be due to multiple different kinds of 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 considered a variety of latent variables that could, either alone or in combination, influence animals' exploratory behavior. We fitted six reinforcement learning (RL) models that could

236 potentially characterize animals' choice behaviors : (1) a "random" model with some overall bias
237 for one choice over the other, (2) a "noisy win-stay lose-shift" model that assumes a win-stay
238 lose shift policy with some level of randomness, (3) a two-parameter "RL" model with a
239 consistent learning rate and some inverse temperature that captures decision noise, (4) a four-
240 parameter "RLCK" model that captures both value-based and value-independent decision with
241 separate parameters for learning rate, decision noise, choice bias, and choice stickiness, (5) a
242 five-parameter "RLCKγ" model that incorporates separate learning rate for rewarded and
243 unrewarded outcomes on top of the "RLCK" model, (6) a five-parameter "RLCKη" model that
244 adds a parameter that tunes the weight between value-based and choice-based decision to the
245 "RLCK" model (**Figure 4A**).

246 First, to identify the model that best captured the computations used by the animals, we
247 compared model fits across six reinforcement learning models with different combinations of
248 latent parameters. AIC weights were calculated from AIC values of each model for each sex and
249 compared across models to determine the best model with the highest relative likelihood.
250 Although model fitting was slightly different across sexes, in both males and females, the
251 "RLCK" model, four-parameter model with value and choice kernel updating policies, best
252 characterized animals' choice behaviors in this task among all six models (**Figure 4B**). Then to
253 confirm that this model was able to capture animals' choice regardless of their learning
254 performance, we used the parameters optimized by the model to simulate choices and compared
255 with the actual choices of animals. We plotted the simulated choices and actual choices against
256 the matching law [23], which dictates that the probability of choice is proportional to the
257 probability of reward. The result supported that this four-parameter model was able to
258 characterize animals' choice behaviors regardless of whether or not the animal was following the

matching law (**Figure 4C**). After comparing all four RL parameters across sexes, we found that females had significantly higher learning rate than males (**Figure 4D**, $t(30) = 2.40$, $p = 0.02$) 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 4E**, 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 were due to differences in learning rate instead of decision noise.

Furthermore, when compared learning rate across sessions within sex, females, but not males, showed increased learning rate over experience with task (**Figure 4F**, RM 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 [24].

Females learned more during exploratory choices than did males.

Our RL modeling revealed that elevated exploration in males was paralleled by elevated learning rates in females. This suggested to us that learning, specifically about gained rewards, may drive differences in explore-exploit strategies. To examine the effect of reward on learning, we conducted win-stay lose-shift analyses across sexes. The result suggested that female mice

281 displayed more win-stay behaviors, indicating that they were more likely than the males to repeat
282 behaviors that produced reward on the previous trial (**Figure 5A**, $t(30) = 2.80$, $p = 0.0089$). As
283 groups, males and females were reasonably, but not perfectly discriminable in terms of the
284 proportion of win stay choices (**Figure 5B,** receiver operating characteristic analysis, AUC =
285 0.74 ± 0.09, 95% CI for the difference = [0.56, 0.92] , $p = 0.0195$). In contrast, male mice tended
286 to shift despite that the previous choice was rewarded. However, there was no significant sex
287 difference in shifting after losses (**Figure 5C**, $t(30) = 0.53$, $p = 0.53$). This suggests that the
288 higher learning rate was mainly driven by win-stay but not lose-shift choices. Then we examined
289 how exploratory state affects reward learning, we calculated the probability of switching on the
290 next trial, given the goal state and reward outcome of the current trial. The result suggested that
291 females were more likely to win stay and lose shift only during exploratory choices, indicating
292 elevated learning during exploration in females (**Figure 5D**, no outlier removed: two-way RM
293 ANOVA, interaction, $F(1,15) = 1.95$, $p = 0.18$; one outlier removed: mix effects analysis, sex X
294 state interaction, $F(1,58) = 4.74$, $p = 0.03$). To account for idiosyncratic strategies (e.g. shifting
295 only after two losses), we used a model-free approach to quantify the extent to which behavior
296 was structured without making strong assumptions about what form this structure might take.

297 To examine how choice behavior was influenced by information of past choice history,
298 given the immediate outcome, we calculated conditioned mutual information for all sessions
299 across sexes [25,26]. The result suggested that mutual information was higher in females than males,
300 suggesting that females were using more information from the past choice and outcome to make
301 their current decision. (**Figure 5E**, $t(30) = 2.65$, $p = 0.013$). As groups, males and females were
302 reasonably discriminable in terms of mutual information (**Figure 5F**, receiver operating
303 characteristic analysis, AUC = 0.74 ± 0.09, 95% CI for the difference = [0.57, 0.92] , p =

0.0195). Together, these results suggested that females were learning more from exploration and utilizing more information from the past trial to make current choices.

Discussion

Sex differences in 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 they 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 change 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 good 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 slower 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 commiting 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 hard 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 [11]. The HMM inferred latent goal states explained more variance in neural activity than other decision-variables [11]. 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 mice models. In 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 (as in **Figure 2C**).

347 Reinforcement learning (RL) models have also been used in the past to identify levels of
348 exploration across individuals [15,16]. However, our findings indicate that multiple latent
349 parameters could concurrently influence how much one explores. While RL models are helpful
350 to understand mechanism questions, they are limited in their ability to identify when exploration
351 is happening, or to characterize individual differences. The HMM model, conversely, has no
352 insight into mechanisms. The HMM and RL models used here could complement each other in
353 understanding the underlying mechanisms of exploration, and how these are influenced across
354 trials and individuals. Since the broader reinforcement learning model framework is highly
355 adaptive and amenable, in the future, the HMM and kinds of model-free analyses we complete
356 here could also inform the design of RL models to more precisely capture explore-exploit
357 decisions.

358 One neuromodulator that is implicated in reinforcement learning, including the transition
359 between exploration and exploitation, *and* is strongly sex-modulated is dopamine [22,27]. Studies
360 have shown that dopamine signal regulates exploration via mechanisms of action selection and
361 learning [22,28,29]. However, due to the exclusive use of one sex in experiments [19,22], the fact that
362 dopamine function on decision making is strongly modulated by sex and sex-linked mechanisms
363 is often overlooked. For example, estradiol has been demonstrated to exert both acute and
364 chronic modulatory effects on dopamine release, reuptake, and receptor binding (Yoest et al.,
365 2014), allowing enhanced DA release and faster DA clearance in females (Becker, 1999). This
366 mechanism could contribute to increased reward learning observed in females during
367 exploration. The prefrontal cortex (PFC), which receives dopamine projections, is a target brain
368 region to understand exploration and exploitation [11,27]. Our previous study implicated PFC in the
369 differences in learning strategy between males and females [5]. It is possible that prefrontal cortical

370 dopamine is particularly engaged in implementing explore-exploit strategies via sex-biased
371 mechanisms of learning.

372 A frequent use of rodent operant testing is to perform cognitive assessments with similar
373 rule structures in animals and dependent measures as are used in human tasks, to permit
374 translational work in animals exploring pharmacology, genetics, and other potential biological
375 contributors to behavior [1,30]. However, the vast majority of these tasks were not designed with
376 computational models in mind, and as a result, we are unable to assess whether similar latent
377 cognitive variables are influencing behavior in humans and rodents. Here, we establish that
378 explore-exploit decisions are shared cognitive processes across species, making this restless
379 bandit task an ideal tool for translational research. Parallel approaches in humans have been used
380 to examine the explore-exploit strategic phenotype of neuropsychiatric disorders, including
381 ADHD, addiction, and depression.[31–34]. The computational modeling used here permits fine
382 grained quantification of individual variability in latent parameters that capture adaptive changes
383 in exploration in changing environments. The new computational tools we develop here could
384 help identify behavioral endophenotypes across species underlying a variety of neuropsychiatric
385 disorders, and open up new avenues for understanding as well as rescuing dysfunction in value-
386 based decision making.

Figures and Figure Captions

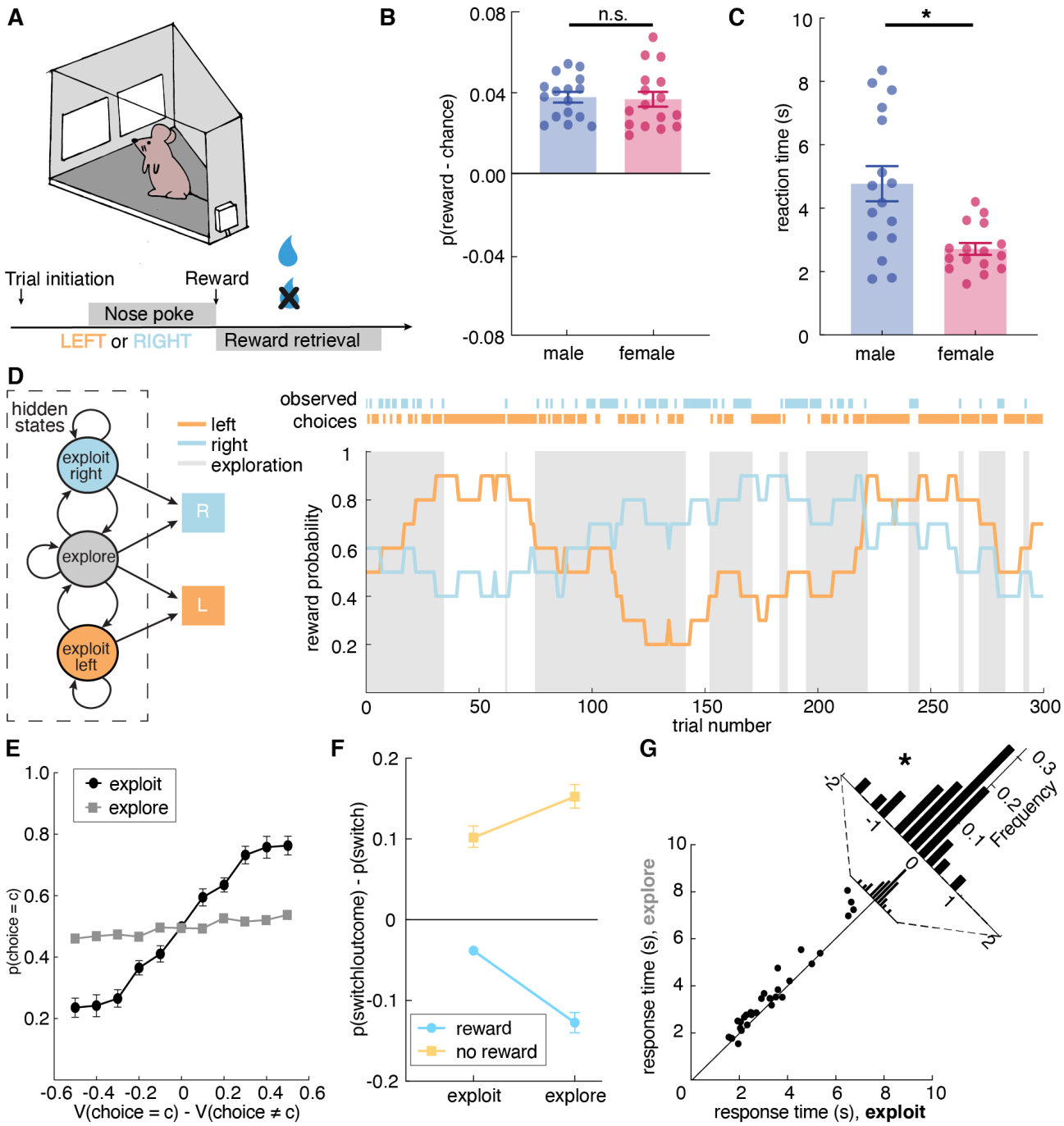

Figure 1. Mice switched between fast switching, outcome-sensitive exploration and slow switching, outcome-insensitive exploitation in a restless bandit task. A) Schematic of the mouse touchscreen chamber with the restless two-armed bandit task and trial structure. B)

Average percentage of reward obtained compared to the chance percentage of reward obtained across sexes. 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 (left and right) 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) Probability of choice as a function of value differences between choices for exploratory and exploitative states. F) The probability of animals switching targets on the next trial, given the current trial's outcome and latent state. G) Difference in choice response time between explore and exploit choices. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

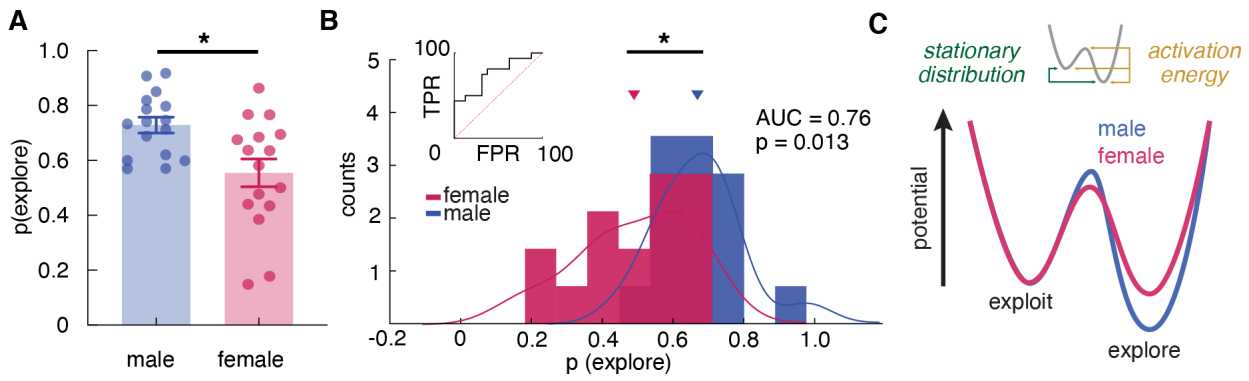

Figure 2. Males explored more than females and they explore for longer periods of time once started. A, B) Average (A) and distribution (B) of percentage of Hidden Markov Model (HMM)-labeled exploratory trials in females and males. C) Dynamic landscape of the fitted HMMs for males and females. The model fit to males had deeper exploratory states, with higher activation energy between the states. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

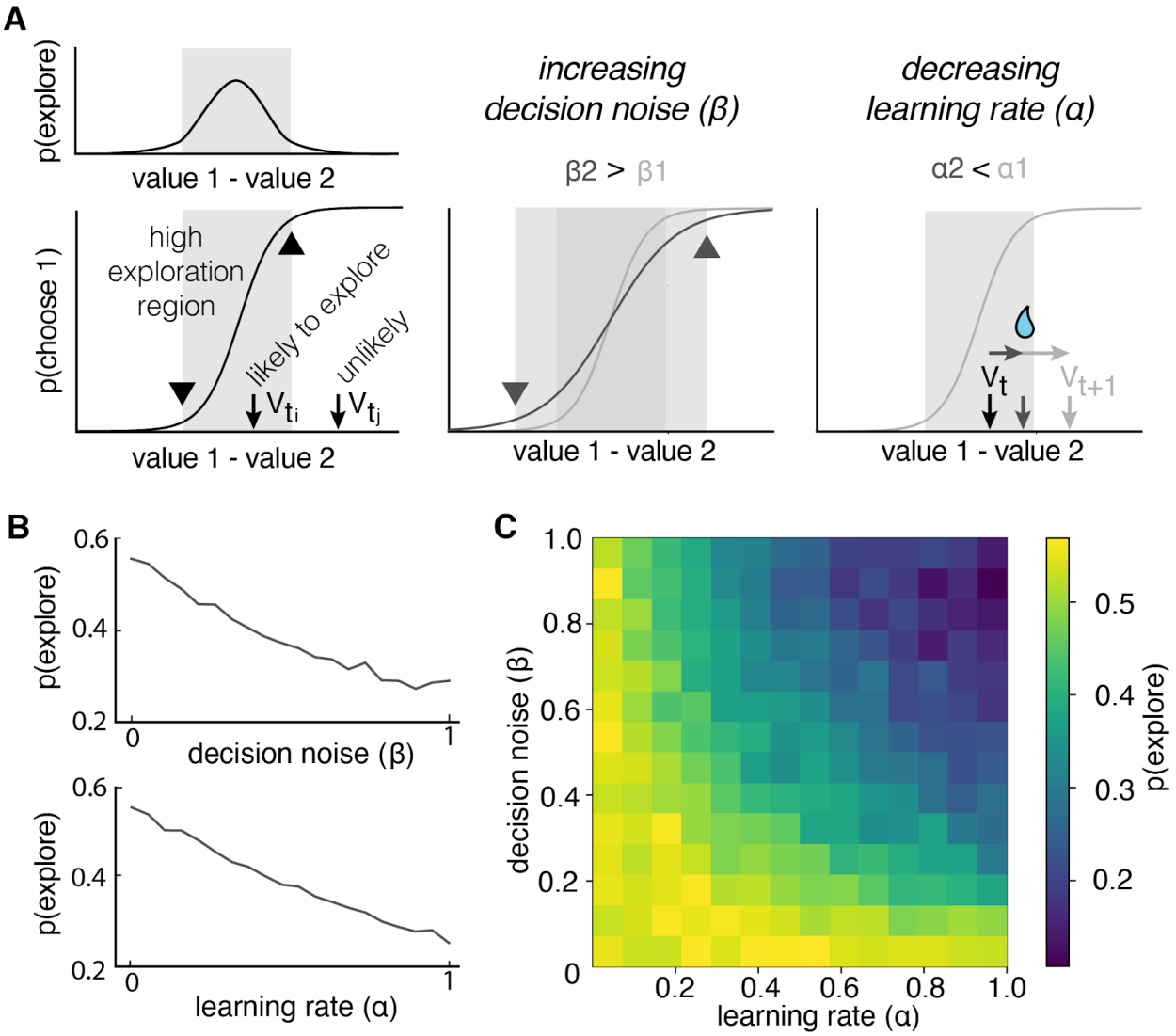

Figure 3. Multiple reinforcement learning parameters can influence the probability of exploration. A) A cartoon illustrating how changing either inverse temperatures (β) or learning rates (α) alters the probability of exploration. As β increases, decision noise decreases and the probability of exploration decreases. As α increases, probability of exploration decreases. B) Computer simulation (10,000 trials) of probability of exploration as a function of either inverse temperatures (β, top) or learning rates (α, bottom). C) Heatmap that demonstrates the level of exploration at different combinations of learning rate and inverse temperature (computer simulation).

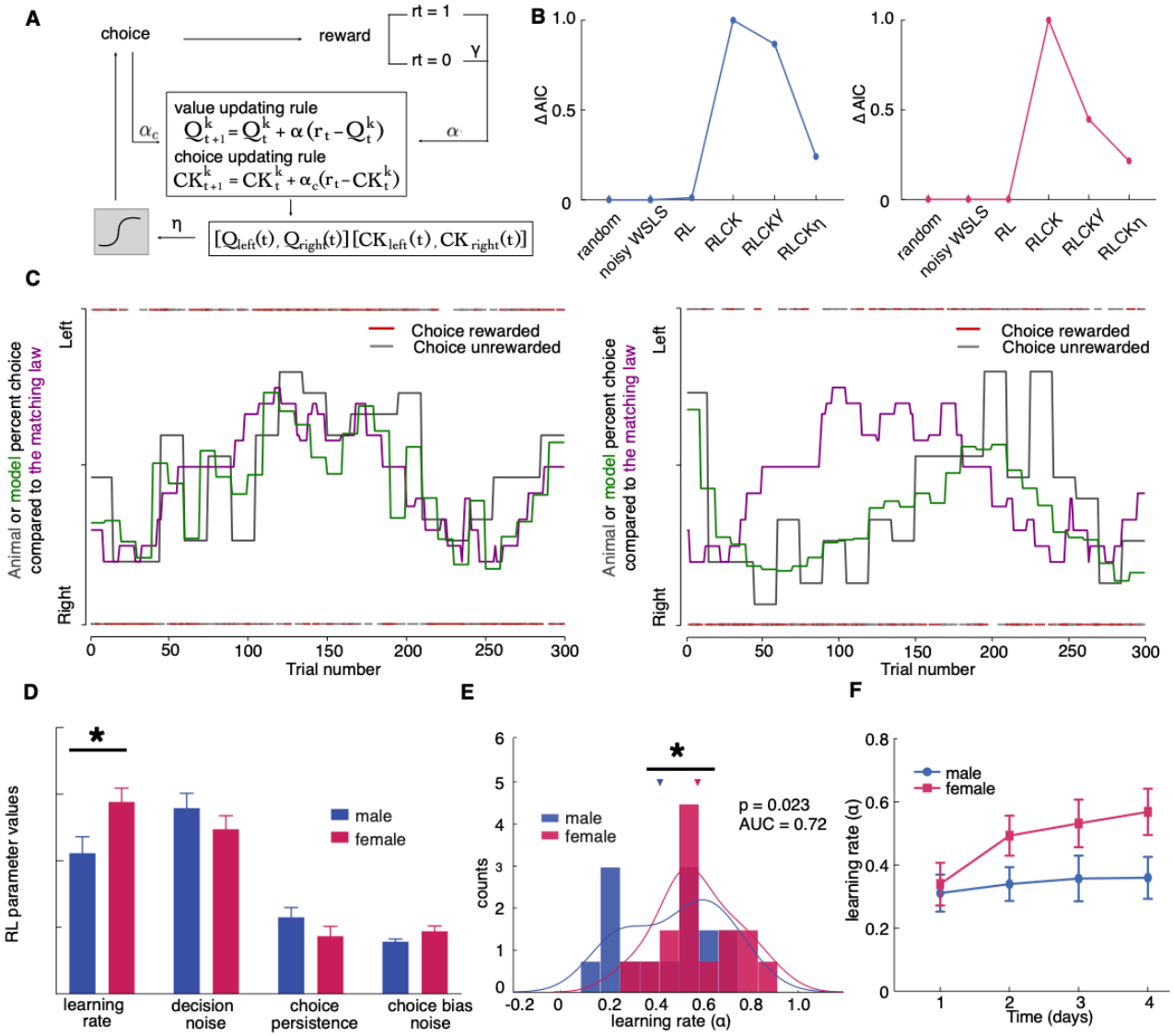

Figure 4. Sex differences in learning rate, but not decision noise, drove differences in explore-exploit decisions. A) A diagram of latent parameters that capture learning, bias, outcomes sensitivity, and noise in reinforcement learning models. The models tested used a combination of these parameters (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) Actual choices (gray) and simulated choices (green) from the best fit model (RLCK)

of two example animals. One animal followed the matching law (purple), and one did not, but in both cases the model was able to capture the actual choice behaviors. D) All four parameters in the best fit RLCK model across sexes. Learning rate (α) was significantly higher in females than males. E) Distribution of learning rate across sexes. F) Learning rate in females increased across days, suggestive of meta learning. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

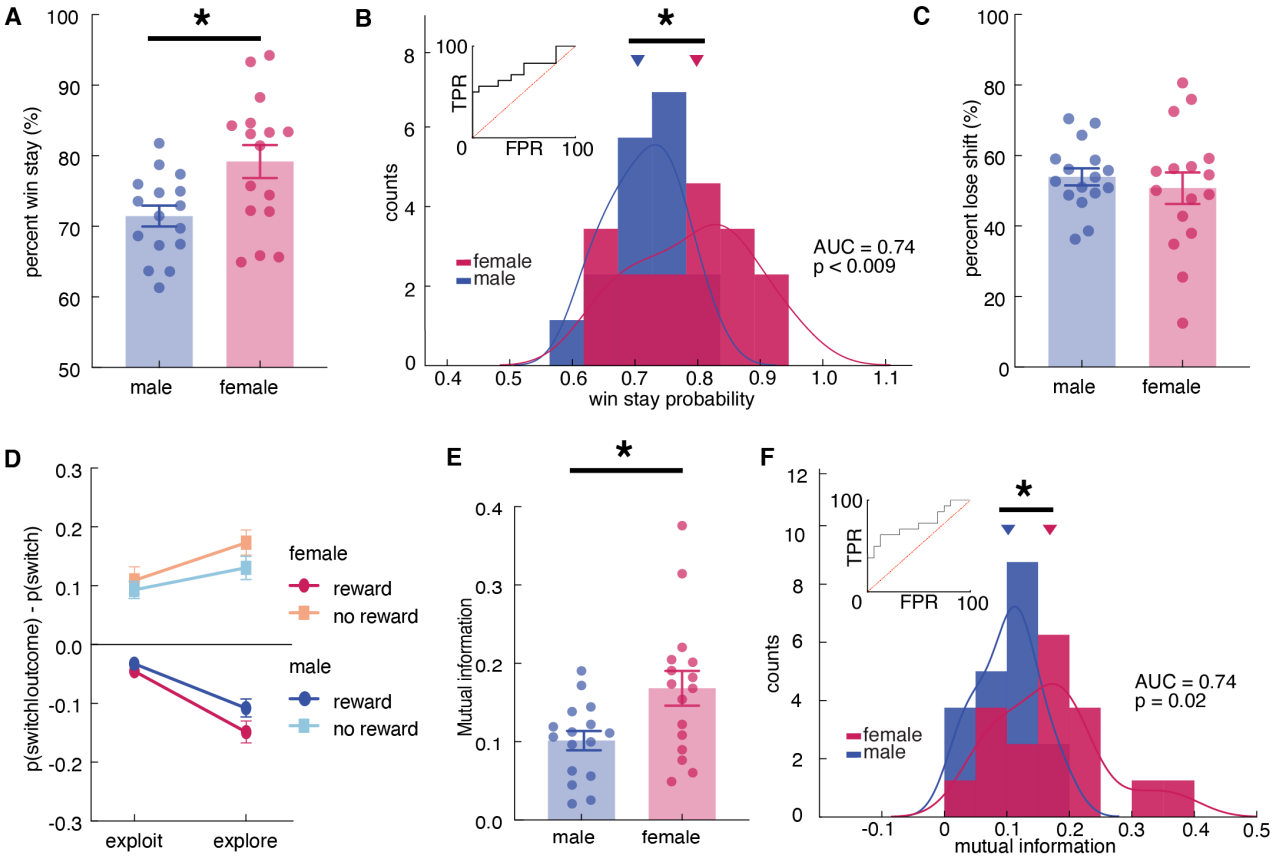

Figure 5. Females used more information about past outcome and past choice to make explore-exploit decisions, and learned more from exploration. A, B) Percent win-stay behavior (A: average; B: distribution) reveals that females were more likely to stay with the same choice after a reward. C) Average percentage of lose-shift behavior across sexes. D) Females were more sensitive to outcomes during exploration than males. The probability of males and females switching targets on the next trial, given the current trial's outcome and latent state. E, F) Average (E) and distribution (F) 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.

445 **Methods**

446 Animals. Thirty-two BL6129SF1/J mice (16 males and 16 females) were obtained from Jackson
447 Laboratories (stock #101043). Mice arrived at the lab at 7 weeks of age, and were housed in
448 groups of four with ad libitum access to water while being mildly food restricted (85-95% of free
449 feeding weight) for the experiment. Animals engaging in operant testing were housed in a 0900–
450 2100 hours reversed light cycle to permit testing during the dark period. Before operant chamber
451 training, animals were food restricted to 85%-90% of free feeding body weight. Operant testing
452 occurred five days per week (Monday-Friday). All animals were cared for according to the
453 guidelines of the National Institution of Health and the University of Minnesota.

454

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

461

462 Behavioral task. T*wo-armed spatial restless bandit task.* Animals were trained to perform a two-
463 armed spatial restless bandit task in the touchscreen operant chamber. Each trial, animals were
464 presented with two identical squares on the left and right side of the screen. Nose poke to one of
465 the target locations on the touchscreen was required to register a response. Each location is
466 associated with some probability of reward, which changes independently over time. For every
467 trial, there is a 10% chance that the reward probability of a given arm will increase or decrease

468 by 10%. All the walks were generated randomly with a few criteria: 1) the overall reward
469 probabilities of two arms are within 2% of each other, preventing one arm being overly better
470 than the other, 2) the reward probability cannot go down to 0% or go up to 100%, 3) there is no
471 30 consecutive trials where the reward probabilities of both arms are lower than 20% to ensure
472 motivation. Animals ran a simple deterministic schedule on Monday to re-adapt to operant
473 chamber after weekends off and ran a different restless bandit task each day from Tuesday to
474 Friday. Animals ran for 2 rounds of 4 consecutive days and within each day, animals completed
475 either 300 trials or spent a maximum of two hours in the operant chamber. Data was recorded by
476 the ABET II system and was exported for further analysis. All computational modeling was
477 conducted using python.
478
479 **Data analysis**
480 General analysis techniques. Data was analyzed with custom PYTHON, MATLAB, and Prism 8
481 scripts. Generalized linear models, ANOVA, and t-test were used to determine sex differences
482 over time, unless otherwise specified. P values were compared against the standard $\alpha = 0.05$
483 threshold. The sample size is $n = 16$ for both males and females for all statistical tests. No animal
484 was excluded from the experiment. One outlier was removed in one analysis using ROUT
485 method (with Q set to 1%). This outlier was from the animal that ran the lowest number of total
486 trials. Statistics for both no outlier removal and outlier removal were reported in the result. All
487 statistical tests used and statistical details were reported in the results. All figures depict mean ±
488 SEM across animals.
489

490 Mixture model. We first asked whether there were different behavioral dynamics that might
491 correspond to exploration and exploitation. Exploration and exploitation take place on different
492 time scales. In RL agents, for example, exploration is typically implemented via adding noise or
493 indeterminacy to a decision-rule. The identity of choices that are caused by this noise—the
494 exploratory choices—will thus switch more frequently than the identity of choices that depend
495 on option value. We should see short runs of exploratory choices and long runs of exploitative
496 ones [11]. To the extent that choice runs end probabilistically (an assumption of the HMM
497 framework), choice run durations (inter-switch intervals) will be exponentially distributed [35]. To
498 the extent that there are multiple causal regimes (such as exploration and exploitation), inter-
499 switch intervals will be distributed as a mixture of multiple exponential distributions
500 (**Supplemental Figure 1**). Because trials are discrete, rather than continuous, we fit mixtures of
501 the discrete equivalent to the exponential distribution, the geometric distribution. We examined
502 the distribution of 24,836 interswitch intervals. Adding a second mixing distribution significantly
503 improved model fit (1-component, 1-parameter mixture log-likelihood: -44555, 2-component, 3-
504 parameter: -41656; likelihood ratio test, $p < 10^{-32}$). Adding additional mixing distributions
505 continued to improve model fit, a common observation in mixture modeling. However, the
506 continued improvement was substantially less than the leap from 1 to 2 components
507 (**Supplemental Figure 1**; 3-component: -41431, 4-component: -41412) and additional mixtures
508 beyond 2 had weights below 3%. This suggests that a mixture of one fast-switching regime and
509 one slow-switching regime was the most parsimonious explanation for the data [36].
510

511 Hidden Markov model. In an HMM framework, choices (y) are "emissions" that are generated
512 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 [8,37]. 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).

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

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 (Blimes Gentle introduction). This algorithm finds a (possibly local) maxima of the complete-data likelihood. A complete set of parameters THETA 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.

<u>Analyzing model dynamics.</u> 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

557 dynamics of exploration, we therefore turned to analytical tools that allowed us to directly
558 characterize the energetic landscape of behavior (**Figure 2C**).
559 In statistical mechanics, processes within a system (like a decision-maker at some
560 moment in time) occupy states (like exploration or exploitation). States have energy associated
561 with them, related to the long-time scale probability of observing a process in those states. A low
562 energy state is one that is very stable and deep, much like a valley between two high peaks. Low
563 energy states will be over-represented in the system. A high energy state, like a ledge on the side
564 of a mountain, is considerably less stable. High energy states will be under-represented in the
565 long-term behavior of the system. The probability of observing a process in a given state i will is
566 related to the energy of that state (E_i) via the Boltzman distribution:

567 $$p_i = \frac{1}{Z} e^{\frac{-E_i}{k_B T}}$$

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

572 $$\frac{p_i}{p_j} = e^{\frac{-(E_i - E_j)}{k_B t}}$$

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

575 $$\ln\left(\frac{p_i}{p_j}\right) k_B T = E_j - E_i$$

576 Meaning that the difference in the energetic depth of the states is proportional to the
577 natural log of the probability of each state, up to some multiplicative factor $k_B T$. To calculate the
578 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 effect 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_BT 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_BT. 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:

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:

$$p_t^L = b \qquad [1] \ \textit{“random”}$$

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:

$$p_t^k = \begin{cases} 1 - \frac{\epsilon}{2}, if\ (c_{t-1} = k\ and\ r_{t-1} = 1\ OR\ c_{t-1} \neq k\ and\ r_{t-1} = 0) \\ \frac{\epsilon}{2}, if\ (c_{t-1} \neq k\ and\ r_{t-1} = 1\ OR\ c_{t-1} = k\ and\ r_{t-1} = 0) \end{cases}$$

[2] *“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.

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

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*:

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

<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).

674 **References**

675 1. Grissom, N. M. & Reyes, T. M. Let's call the whole thing off: evaluating gender and sex
676 differences in executive function. *Neuropsychopharmacology* (2018) doi:10.1038/s41386-
677 018-0179-5.
678 2. Green, T., Flash, S. & Reiss, A. L. Sex differences in psychiatric disorders: what we can
679 learn from sex chromosome aneuploidies. *Neuropsychopharmacology* **44**, 9–21 (2019).
680 3. Shansky, R. M. Are hormones a 'female problem' for animal research? *Science* **364**, 825–
681 826 (2019).
682 4. Maney, D. L. Perils and pitfalls of reporting sex differences. *Philos. Trans. R. Soc. Lond. B*
683 *Biol. Sci.* **371**, 20150119 (2016).
684 5. Chen, C. S. *et al.* Divergent Strategies for Learning in Males and Females. *Curr. Biol.* **0**,
685 (2020).
686 6. Wilson, R. C., Bonawitz, E., Costa, V. D. & Ebitz, R. B. Balancing exploration and
687 exploitation with information and randomization. *Curr Opin Behav Sci* **38**, 49–56 (2021).
688 7. Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L. & Platt, M. L. A Primer on
689 Foraging and the Explore/Exploit Trade-Off for Psychiatry Research.
690 *Neuropsychopharmacology* **42**, 1931–1939 (2017).
691 8. Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W. & Hayden, B. Y. Tonic
692 exploration governs both flexibility and lapses. *PLoS Comput. Biol.* **15**, e1007475 (2019).
693 9. Pisupati, S., Chartarifsky-Lynn, L., Khanal, A. & Churchland, A. K. Lapses in perceptual
694 judgments reflect exploration. *Cold Spring Harbor Laboratory* 613828 (2019)
695 doi:10.1101/613828.
696 10. van den Bos, R., Homberg, J. & de Visser, L. A critical review of sex differences in

decision-making tasks: Focus on the Iowa Gambling Task. *Behavioural Brain Research* vol. 238 95–108 (2013).

11. Ebitz, R. B., Albarran, E. & Moore, T. Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. *Neuron* **97**, 475 (2018).
12. Filipowicz, A. L. S. *et al.* The complexity of model-free and model-based learning strategies. (2019) doi:10.1101/2019.12.28.879965.
13. Kurdi, B., Gershman, S. J. & Banaji, M. R. 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 (2019).
14. Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. Decision making and the avoidance of cognitive demand. *J. Exp. Psychol. Gen.* **139**, 665–682 (2010).
15. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
16. Pearson, J. M., Hayden, B. Y., Raghavachari, S. & Platt, M. L. Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Curr. Biol.* **19**, 1532–1537 (2009).
17. Jepma, M. & Nieuwenhuis, S. Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory. *J. Cogn. Neurosci.* **23**, 1587–1596 (2011).
18. Ishii, S., Yoshida, W. & Yoshimoto, J. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Netw.* **15**, 665–687 (2002).
19. Cinotti, F. *et al.* Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Sci. Rep.* **9**, 6770 (2019).

720 20. Schweighofer, N. & Doya, K. Meta-learning in reinforcement learning. *Neural Netw.* **16**, 5–
721 9 (2003).
722 21. Averbeck, B. *et al.* 376. Subcortical Contributions to the Explore-Exploit Tradeoff.
723 *Biological Psychiatry* vol. 81 S154 (2017).
724 22. Beeler, J. A., Daw, N., Frazier, C. R. M. & Zhuang, X. Tonic dopamine modulates
725 exploitation of reward learning. *Front. Behav. Neurosci.* **4**, 170 (2010).
726 23. Poling, A., Edwards, T. L., Weeden, M. & Foster, T. M. The Matching Law. *Psychol. Rec.*
727 **61**, 313–322 (2011).
728 24. Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15**, 495–506 (2002).
729 25. Leao, D., Jr, Fragoso, M. & Ruffino, P. Regular conditional probability, disintegration of
730 probability and Radon spaces. *Proyecciones* **23**, 15–29 (2004).
731 26. Wyner, A. D. A definition of conditional mutual information for arbitrary ensembles.
732 *Information and Control* **38**, 51–59 (1978).
733 27. Jenni, N. L., Larkin, J. D. & Floresco, S. B. Prefrontal Dopamine D1 and D2 Receptors
734 Regulate Dissociable Aspects of Decision Making via Distinct Ventral Striatal and
735 Amygdalar Circuits. *J. Neurosci.* **37**, 6200–6213 (2017).
736 28. Frank, M. J., Doll, B. B., Oas-Terpstra, J. & Moreno, F. Prefrontal and striatal dopaminergic
737 genes predict individual differences in exploration and exploitation. *Nat. Neurosci.* **12**,
738 1062–1068 (2009).
739 29. Humphries, M. D., Khamassi, M. & Gurney, K. Dopaminergic Control of the Exploration-
740 Exploitation Trade-Off via the Basal Ganglia. *Front. Neurosci.* **6**, 9 (2012).
741 30. Heath, C. J., Phillips, B. U., Bussey, T. J. & Saksida, L. M. Measuring Motivation and
742 Reward-Related Decision Making in the Rodent Operant Touchscreen System. *Curr.*

Protoc. Neurosci. **74**, 8.34.1–20 (2016).

31. Addicott, M. A. *et al.* Attention-deficit/hyperactivity disorder and the explore/exploit trade-off. *Neuropsychopharmacology* (2020) doi:10.1038/s41386-020-00881-8.
32. Blanco, N. J., Otto, A. R., Maddox, W. T., Beevers, C. G. & Love, B. C. The influence of depression symptoms on exploratory decision-making. *Cognition* **129**, 563–568 (2013).
33. Addicott, M. A. *et al.* Smoking withdrawal is associated with increases in brain activation during decision making and reward anticipation: a preliminary study. *Psychopharmacology* **219**, 563–573 (2012).
34. Beeler, J. A., Frazier, C. R. M. & Zhuang, X. Putting desire on a budget: dopamine and energy expenditure, reconciling reward and resources. *Front. Integr. Neurosci.* **6**, 49 (2012).
35. Berg, H. C. *Random Walks in Biology*. (Princeton University Press, 1993).
36. McLachlan, G. J. & Peel, D. *Finite Mixture Models*. (John Wiley & Sons, 2004).
37. Ebitz, R. B., Tu, J. C. & Hayden, B. Y. Rules warp feature encoding in decision-making circuits. *PLoS Biol.* **18**, e3000951 (2020).