Dopamine and norepinephrine differentially mediate the exploration-exploitation tradeoff

Cathy S. Chen[1], Dana Mueller[1], Evan Knep[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

The catecholamines dopamine (DA) and norepinephrine (NE) have been repeatedly implicated in neuropsychiatric vulnerability, in part via their roles in mediating the decision making processes. Although the two neuromodulators share a synthesis pathway and are co-activated under states of arousal, they engage in distinct circuits and roles in modulating neural activity across the brain. However, in the computational neuroscience literature, they have been assigned similar roles in modulating the latent cognitive processes of decision making, in particular the exploration-exploitation tradeoff. Revealing how each neuromodulator contributes to this explore-exploit process will be important in guiding mechanistic hypotheses emerging from computational psychiatric approaches. To understand the differences and overlaps of the roles of these two catecholamine systems in regulating exploration and exploitation, a direct comparison using the same dynamic decision making task is needed. Here, we ran mice in a restless two-armed bandit task, which encourages both exploration and exploitation. We systemically administered a nonselective DA receptor antagonist (flupenthixol), a nonselective DA receptor agonist (apomorphine), a NE beta-receptor antagonist (propranolol), and a NE beta-receptor agonist (isoproterenol), and examined changes in exploration within subjects across sessions. We found a bidirectional modulatory effect of dopamine receptor activity on the level of exploration. Increasing dopamine activity decreased exploration and decreasing dopamine activity increased exploration. Beta-noradrenergic receptor activity also modulated exploration, but the modulatory effect was mediated by sex. Reinforcement learning model parameters suggested that dopamine modulation affected exploration via decision noise and norepinephrine modulation affected exploration via outcome sensitivity. Together, these findings suggested that the mechanisms that govern the transition between exploration and exploitation are sensitive to changes in both catecholamine functions and revealed differential roles for NE and DA in mediating exploration.

Introduction

Dysfunctions in cognitive processes, particularly within the domain of executive function, are implicated in numerous neuropsychiatric disorders [1–3]. One essential aspect of executive function is value-based decision making. Decision making involving exploring and learning in a dynamic environment has been a focus in computational neuroscience, which has developed approaches for discovering latent structures in decision making strategies [4–7]. Many computational models for value-based decision making distinguish two essential latent processes: exploration and exploitation [8–12]. Exploration is a ubiquitous learning process across species that allows the discovery of the best action to take in an uncertain environment [6,9,13]. Once a favorable action or option is discovered, exploitation, i.e. repeating a rewarding action, is necessary to obtain the best rates of reward. However, exploitation must be balanced with continued exploration as environments and their reward probabilities change. Dysregulation in this exploration-exploitation tradeoff can be observed in the phenotypes of numerous neuropsychiatric disorders and challenges, including schizophrenia, autism spectrum disorders, addictions, and chronic stress [14–18]. Understanding the fundamental mechanisms impacting the computations in our brain that underlie the balance between exploration and exploitation could help identify critical circuits that are associated with differential risk factors of neuropsychiatric disorders and open avenues for novel interventions for executive function challenges.

Many neuropsychiatric challenges are associated with dysfunction in the catecholamines dopamine and norepinephrine [16,19–24]. These neuromodulators are well-positioned to carry information about the state of the environment and influence behavior outputs via downstream control of action selection [9,25]. Indeed, dopamine and norepinephrine have each separately been implicated in mediating the exploration-exploitation tradeoff by promoting exploration [9,23–28]. While dopamine and norepinephrine share biosynthetic pathways and are co-activated under states of arousal, they act through partially separable circuits and have distinct pharmacological profiles [29]. It is notable, therefore, that dopamine and norepinephrine have largely been ascribed similar or even identical roles in mediating the explore-exploit tradeoff in the computational neuroscience literature.

We have previously shown that multiple latent cognitive processes, such as the reinforcement learning model-derived parameters of learning rate and decision noise, can describe differences in the rates of exploration between groups [4]. Dopamine has often been ascribed a role in mediating reward prediction errors and value-based learning [10,30–32] while norepinephrine in modulating arousal, attention, and value assessment [23,33,34], but more recently both neuromodulators have been implicated in the process of action selection [35,36]. Therefore, it is unknown whether dopamine and norepinephrine govern this exploration-exploitation tradeoff via distinct or similar latent cognitive processes because they have been largely studied in this context in isolation [26].

To uncover the distinct roles of dopamine and norepinephrine in the exploration-exploitation tradeoff, we conducted within-subjects pharmacological manipulations for these two neuromodulators on exploration, and the latent cognitive parameters that influence exploration, in the same restless bandit task. We pharmacologically up- and down-regulated dopaminergic and noradrenergic receptor activity and compared the modulatory effects on exploration in the same dynamic decision making task, a spatial restless two-armed bandit [4,6]. We found a bidirectional modulatory effect of dopamine receptor activity on the level of exploration. Increasing dopamine activity decreased exploration and decreasing dopamine activity increased exploration. Beta-noradrenergic receptor activity also modulated exploration but the modulatory effect was mediated by sex. Contrary to dopamine modulating choice solely by changing value-based learning for outcomes, via reinforcement learning modeling we find that dopamine mediates exploration by changing decision noise, i.e. the precision of value-based choice selection. In contrast, noradrenergic activity surprisingly influenced exploration by changing value-based learning for outcomes in a sex dependent manner. These data suggest differential roles of dopamine and norepinephrine in mediating exploration, and complex roles for both catecholamines in signaling reward.

Results

Aged-matched adult male and female wildtype mice (n = 32, 16 males and 16 females, strain B6129SF1/J) were trained to perform a restless two-armed spatial bandit task in touch-screen operant chambers (**Figure 1A**). In this task, animals were presented with two identical visual targets (squares) on the left and right side of the screen during each trial. They indicated their choices by nose poking at one of the two target locations, each of which provided some probability of reward that changed independently and randomly over time (**Figure 1B)**. Animals had to infer which location offered better payoff by constantly sampling the two locations and learning from the reward outcomes. The dynamic reward contingency of this task naturally encourages the animal to balance between exploiting a favorable option when it is found and exploring to gain information about potential better alternatives. This task has been employed in rodents and primates as well as human subjects to understand learning and exploration and have successfully revealed divergent exploration strategies [4,6,37]. In this study, we adopted this task to understand the modulatory effect of catecholamine receptor activity on the transition between exploration and exploitation.

To examine the distinct and/or overlapping roles of dopamine and norepinephrine in modulating exploration, we tested the systemic effect of a non-selective dopamine receptor agonist apomorphine (0.1mg/kg), a non-selective dopamine receptor antagonist flupenthixol (0.03mg/kg), a beta-noradrenergic receptor agonist isoproterenol (0.3mg/kg), and a beta-noradrenergic receptor antagonist propranolol (5mg/kg) on exploration. The dosages of the drugs were based on previous studies on the role of dopamine and norepinephrine in cognitive processes [38–44]. Animals were intraperitoneally (IP) administered either saline (control) or treatment with 5 mg/kg of propranolol, 0.3mg/kg of isoproterenol, or 0.03 mg/kg of flupenthixol prior to the bandit task or 0.1mg/kg of apomorphine 30 mins prior to the bandit task. Saline and drug treatments were administered interleaved by session (**Figure 1C, 2A**). Mice performed 12 sessions under each pharmacological condition (6 control and 6 drug treatment sessions). The experiment used a within-subject design, where animals received each of the four drugs tested, allowing us to directly compare the effect of drugs within individual animals. Each session consisted of 300 trials.

To examine whether mice had learned to perform the task and were able to sustain reward, we first calculated the average probability of obtaining reward across sessions in all control and treatment groups. It's worth noting that in this task, there is no asymptotic performance because the animals need to constantly adapt to the changing reward contingency and any one option is not always the best choice to make. At any given time, an animal could decide to repeat a favorable option or to explore the options. The performance is best measured by the amount of reward acquired. Because reward schedules were dynamic and stochastic, sessions could differ slightly in the amount of reward that was available. We therefore compared the probability of actual obtained reward against the probability of reward if chosen randomly (chance). In the dopamine modulation condition, both control groups (saline) and treatment groups (apomorphine and flupenthixol) were able to obtain reward more frequently than chance (**Figure 1D, E**, one-sample t test, vehicle control (apomorphine): $t(31) = 13.96$, $p<0.0001$; apomorphine: $t(31)= 17.24$, $p<0.0001$; vehicle control (flupenthixol): $t(31)=12.36$, $p<0.0001$; flupenthixol: $t(31)=16.48$, $p<0.0001$). When modulating beta-noradrenergic receptor activity, both control groups (saline) and treatment groups (isoproterenol and propranolol) were able to obtain reward more frequently than chance (**Figure 2B, D**, one-sample t test, vehicle control (isoproterenol): $t(31)=10.01$, $p<0.0001$; isoproterenol: $t(31)=13.84$, $p<0.0001$; vehicle control (propranolol): $t(31)=10.60$, $p<0.0001$; propranolol: $t(31)=20.08$, $p<0.0001$). These results suggested that animals were not choosing randomly. The animals demonstrated a strong understanding of the task and were able to effectively obtain rewards above chance under baseline (vehicle control) and all drug treatments.

Next, we asked whether manipulating dopaminergic receptor activity influenced their reward acquisition performance. We used a generalized linear mixed model (GLMM) to estimate effects and interactions of drug treatments and sex, with session and animal identity as random effects (**Equation 1**, see methods). Increasing dopamine receptor activity with apomorphine administration did not significantly alter their reward acquisition performance (**Figure 1D**, GLMM, main effect of drug, $p= 0.11$, $\beta1 = -0.017$ **)**. Antagonizing dopamine receptor activity with flupenthixol administration also did not significantly alter reward acquisition performance (**Figure 1E**, GLMM, main effect of drug, $p= 0.28$, $\beta1 = -0.008$). Increasing beta-noradrenergic

receptor activity with isoproterenol administration did not affect the amount of reward obtained (**Figure 2B**, GLMM, main effect of drug, $p= 0.93$, $\beta1 = -0.0005$). However, when decreasing beta-noradrenergic receptor activity with propranolol administration, animals obtained significantly more reward (**Figure 2D**, GLMM, main effect of drug, $p < 0.0001$, $\beta1 = -0.497$). Even though manipulating dopamine receptor activity and increasing beta-noradrenergic receptor activity did not affect how much overall obtained reward, we considered the possibility that the tuning of catecholamine receptor activity affect latent exploration strategies underlying choice patterns, but not an observable change in the level of accuracy. Our previous study demonstrated that similar reward acquisition performance can be achieved via divergent explore strategies [4].

Pharmacological manipulations of both dopaminergic and noradrenergic receptor activity have been shown to have a dose-dependent effect on motor functions [44,45]. In this study, dosage selection was based on previous literature examining cognitive functions and a low dose was used to have influence on cognition but avoid motor function impairment [38,41]. One commonly used behavioral metric to examine motor function in a cognitive task is response time. Therefore, we calculated the response time, which was the time elapsed between the onset of choices and nose poke response to make a choice, as recorded by the touchscreen operant chamber. If the administration of drugs impaired motor function, we would expect to see longer response time in the drug group compared to vehicle control. Upregulating dopamine receptor activity with apomorphine did not significantly influence the response time (GLMM, main effect of drug, $p=0.77$, $\beta1 = 0.10$). Similarly, downregulating dopamine receptor activity with flupenthixol did not significantly alter animals' response time (GLMM, main effect of drug, $p=0.96$, $\beta1 = -0.01$). Modulating beta-noradrenergic receptor activity resulted in bidirectional changes in response time. When increasing beta-noradrenergic receptor activity with isoproterenol, both males and females took significantly longer to respond (**Figure 2C,** GLMM, main effect of drug, $p<0.0001$, $\beta1 = -3.51$). Decreasing beta-noradrenergic receptor activity with propranolol significantly decreased the response time (**Figure 2E,** GLMM, main effect of drug, $p=0.04$, $\beta1 = 0.51$), primarily driven by response time reduction in females under propranolol compared to vehicle controls (GLMM, interaction term, $p<0.009$, $\beta3 = -0.57$). Faster response time under propranolol could be associated with more reward obtained - propranolol was the only drug treatment that altered reward acquisition performance among the four drugs tested.

Previous studies have linked response time with the complexity of the strategy used [4,46,47]. Faster response time under propranolol might also point to the possibility that decreasing beta-noradrenergic receptor activity brought on changes in strategies that took shorter time to execute. Isoproterenol was a highly significant predictor for lower response time in both male and female mice, implicating the potential drug effect on motor behaviors.

Opposing modulatory effect of dopamine and norepinephrine receptor activity on win-stay lose-shift behaviors

The role of dopamine has been heavily studied and shown to be a key contributor to value-based decision making and reinforcement learning [27,28,31,48]. Although modulating dopamine receptor activity did not significantly influence the overall amount of reward obtained or the response time, it is possible that apomorphine and flupenthixol exerted influence on how animals adapted their choices to reward outcome. To understand how increasing and decreasing dopamine activity influenced animals' reward-driven choice behaviors, we examined the probability of win-stay (repeating a choice when it was rewarded on the previous trial) and the probability of lose-shift (switching to the other choice when not rewarded). We found that when increasing dopamine receptor activity with apomorphine administration, both males and females had increased win-stay behavior (**Figure 1F,** GLMM, main effect of drug, $p=0.011$, $\beta1= -0.02$) and decreased lose-shift behavior (**Figure 1G,** GLMM, main effect of drug, $p=0.016$, $\beta1= 0.034$). This result suggested that apomorphine increased the overall stickiness of choice and more stay behaviors regardless of the outcome.

There was no significant change in the probability of win-stay when the animals were administered flupenthixol (**Figure 1H,** GLMM, main effect of drug, $p=0.27$, $\beta1= -0.001$). It's worth noting that there was baseline sex difference in the probability of win-stay, with females having a higher probability of win-stay (GLMM, main effect of sex, apomorphine: $p=0.065$, $\beta2 = -0.054$; flupenthixol: $p=0.01$, $\beta2 =-0.05$). The results also revealed an opposing effect from dopamine agonist that antagonizing dopamine receptor activity with flupenthixol administration increased the probability of lose-shift (**Figure 1I,** GLMM, main effect of drug, $p=0.04$, $\beta1 = -0.014$).

To control for the total amount of switching, we calculated outcome sensitivity (**Equation 2,** see Methods) to understand how much of the switching behavior was outcome-sensitive. The result suggests that there were no significant changes in outcome sensitivity when animals were administered apomorphine (GLMM, main effect of drug, p=0.22, β1 = -0.08). However, administration of flupenthixol significantly increased outcome sensitivity across both sexes (**Figure S1A,** GLMM, main effect of drug, p=0.006, β1= -0.06) and there was more increase in outcome sensitivity in males (GLMM, interaction term, p=0.027, β3= -0.10).

Modulating beta-noradrenergic receptor activity also influenced win-stay and lose-shift behaviors. Increasing beta-noradrenergic activity with isoproterenol administration decreased the probability of win-stay compared to vehicle control (**Figure 2F,** GLMM, main effect of drug, 0.009, β1=0.007). Isoproterenol administration also decreased the probability of lose-shift in both sexes (**Figure 2G,** GLMM, main effect of drug, p=0.0007, β1=0.083), primarily driven by significant decrease in lose-shift in females under isoproterenol (**Figure 2G inset,** GLMM, interaction term, p=0.004, β3=-0.077). Outcome sensitivity analysis revealed a significant decreased in outcome sensitivity when administer isoproterenol compared to vehicle control (**Figure S1B,** GLMM, main effect of drug, p=0.021, β1 = 0.134). In contrast, decreasing beta-noradrenergic activity with propranolol administration increased the probability of win-stay compared to vehicle control (**Figure 2H,** GLMM, main effect of drug, p<0.0001, β1 = -0.023). There was also a significant interaction between sex and drug that males showed a greater increase in win-stay on propranolol compared to females (**Figure 2H inset,** GLMM, interaction term, 0.02, β3=-0.036). Interestingly, propranolol administration also decreased the probability of lose-shift but only in males (**Figure 2I,** GLMM, main effect of drug, 0.45, β1 = -0.01, interaction term, p= 0.002, β3 =0.08). Outcome sensitivity analysis revealed a significant increase in outcome sensitivity with propranolol administration (**Figure S1C,** GLMM, main effect of drug, 0.003, β1 = -0.186). Together, these results suggested an opposing role of dopamine and norepinephrine in modulating reward-driven decision making. Increasing dopamine or decreasing norepinephrine activity resulted in increased win-stay and decreased lose-shift but this effect is more sex-different in the noradrenergic manipulations.

A Hidden Markov model (HMM) identifies distinct patterns of exploration and exploitation associated with modulation of dopamine and norepinephrine activity.

The changes in the reward-driven behaviors such as win-stay and lose-shift could be a manifestation of strategic changes in how animals explored the changing environment. To understand how modulation of dopamine and norepinephrine affect the balance between exploration and exploitation, we used a Hidden Markov model (HMM) that modeled exploration and exploitation as two latent goal states to infer which choices were exploratory or exploitative (**Figure 3A)**. Figure 3A showed an example of HMM labeling of choices of an animal in the restless bandit task. The shaded area indicates exploratory trials inferred by the HMM. The animal displayed mixtures of exploratory bouts where choices were distributed across two choices and exploitative bouts where one choice was repeated. The HMM model has previously been used in rodents, primates and humans [4,6,49] to infer explore-exploit state from choices and has shown robust correlation with neural activity as well as other behavioral metrics, including response time, value function, and reinforcement learning.

Essentially, when a HMM model is fitted to animals' choices, it estimates a transition matrix, which is the mapping between every state at one time point and every state at some future point of time. The HMM model fitting is the process of estimating how the distribution over states will evolve over time. The transition matrix estimates the probability of transition between states - that is the probability of staying in the explore state, the probability of transitioning from explore state to exploit state, the probability of transitioning from exploit state to explore state, and the probability of staying in exploit state. Note that exploration is modeled as a uniform distribution over choices. This is because the uniform distribution over choices is the maximum entropy distribution of categorical variables. We want to make as few assumptions as we possibly can about the choices they make during exploration. With the HMM transition matrix optimized from each animal's choice sequences, we then used the Viterbi algorithm to decode latent states from choices, allowing us to label each choice as either exploratory or exploitative.

To examine how much animals were exploring, we calculated the average number of exploratory choices inferred from HMM. The results revealed a bidirectional modulatory effect of dopamine

285 receptor activity on the level of exploration. When animals were administered apomorphine, we
286 found that animals on average had fewer exploratory trials than when they were on vehicle
287 control, with 55.5% ± 12.6% STD of trials labeled as exploratory on apomorphine and 72.3% ±
288 13.6% STD of trials labeled as exploratory on vehicle (**Figure 3B, D,** GLMM, main effect of
289 drug, $p<0.0001$, $\beta1 = 0.174$). This is also consistent with our findings in the win-stay lose-shift
290 analysis that apomorphine increased the stickiness or repetitiveness of choice behaviors. In
291 contrast, when decreasing dopamine receptor activity with flupenthixol administration, animals
292 explore more compared to vehicle control, with 58.2% ± 12.0% STD of trials being exploratory
293 on flupenthixol and 52.1% ± 12.3% STD of trials being exploratory on vehicle control (**Figure**
294 **3B, D**, GLMM, main effect of drug, 0.002, $\beta1 = -0.062$). This result suggested that manipulating
295 dopamine receptor activity bidirectionally affected the level of exploration across both sexes
296 (**Figure 3C, Figure S2)** - increasing dopamine receptor activity decreased exploratory choices
297 and decreasing dopamine receptor activity increased exploratory choices.
298
299 Modulating beta-noradrenergic receptor activity also influenced the probability of exploration
300 and the modulatory effect on exploration was associated with sex. Decreasing beta-noradrenergic
301 receptor activity with propranolol significantly decreased the level of exploration (**Figure 4A, C,**
302 **D**) but only in males (**Figure 4F, Figure S2**, GLMM, interaction term, 0.0001, $\beta1 = 0.138$).
303 Increasing beta-noradrenergic receptor activity with isoproterenol administration significantly
304 decreased exploration in females (**Figure 3G, 3E, Figure S2**, interaction term, $p= 0.002$, $\beta3=-$
305 0.103). Together these results suggested a sex-linked neuromodulatory effect of beta-
306 noradrenergic receptor on exploration - increasing beta-noradrenergic activity decreased
307 exploration in males and decreasing beta-noradrenergic activity decreased exploration in
308 females. One possibility is that this sex-differentiated modulatory effect reflected sex-dependent
309 ceiling/floor effects of noradrenergic signaling. Another possibility is that there is a potential
310 inverted U-shape relationship between beta-noradrenergic receptor activity level and exploration
311 (**Figure 4H**). Previous studies have suggested higher tonic norepinephrine (NE) levels in females
312 due to estrogen-mediated enhanced NE synthesis and reduced NE clearance [50]. The difference
313 in baseline NE activity across sexes may contribute to the differential modulatory effect on
314 exploration when increasing or decreasing in noradrenergic receptor activity.
315

Reinforcement learning (RL) models revealed changes in distinct parameters under modulation of dopamine and epinephrine

The results of the HMM suggested that both dopamine and norepinephrine modulation influenced the level of exploration. However, it is unclear whether changes in exploration were due to changes in similar or distinct latent cognitive parameters. Does dopamine and norepinephrine modulate exploration via distinct or overlapping mechanisms? To address this question, we fitted a series of reinforcement learning (RL) models to understand the effect of pharmacological manipulation on the latent cognitive parameters that could influence exploration and exploitation [10,51–53].

To make inferences about changes in the RL model parameters, we first identified the best-fitting RL model for the animals' behaviors. The majority of RL model parameters can be categorized in three ways: value-dependent learning terms, value-independent bias terms, and decision noise/randomness terms [54]. Previous studies demonstrated the effect of various RL parameters on value-based decision making, including value-dependent terms such as learning rate and asymmetrical learning rate [4,38,55,56], value-independent terms such as choice bias [4,54,57], and noise terms such as inverse temperature and lapse rate [38,57,58]. Here, we compared six reinforcement learning models that incorporated one or more of the above parameters (learning rate, bias, noise) and had different assumptions about the latent cognitive processes that animals might employ during exploration (**Figure 5A, Equations 3-7)**. These models included: (1) a "random" model with some overall bias for one choice over the other, (2) a "noisy win-stay lose-shift" model (WSLS) 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 three-parameter "RLCK" model that captures both value-based and value-independent decision with two parameters for learning rate and choice bias, and an overall decision noise parameter, (5) a three-parameter "RLγ" model that captures asymmetrical learning with a learning rate parameter and a scaling parameter for negative outcome learning and a decision noise parameter, and (6) a four-parameter "RLCKγ" asymmetrical learning and bias model that includes a choice bias term on top of the "RLγ" (see methods). We fitted these

346 models to each individual animal and compared the likelihood of each of the six models under
347 four vehicle control conditions and four drug conditions.
348
349 In dopamine modulation conditions, the RLCK model (learning + choice kernel,) and the
350 RLCKγ (asymmetrical learning and bias) model best fit the behaviors (**Figure 5B,** average AIC
351 across DA conditions: model 1 random: 73838.03; model 2 WSLS: 55843.26; model 3 RL:
352 58133.45; model 4 RLCK: 54652.02; model 5 RLγ: 57545.47; model 6 RLCKγ: 54612.56).
353 Since the RLCKγ model did not significantly improve the model fitting, we decided to use the
354 more parsimonious RLCK model that has three parameters (learning rate, choice bias, and
355 decision noise). Then we examined how the reinforcement learning model parameters changed
356 with dopamine modulation. Previous study has shown that both learning rate and inverse
357 temperature parameters could induce changes in the overall level of exploration [4]. Therefore,
358 we asked whether the bidirectional changes in exploration when up- or down-regulating
359 dopamine receptor activity were due to changes in learning rate or decision noise. The results
360 revealed a bidirectional effect of dopamine modulation on the inverse temperature parameter
361 (**Figure 5C-E**). Because RL model data can be non-normally distributed, the Shapiro-Wilk test
362 of normality was conducted to determine whether the RL parameters were normally distributed.
363 The result suggested that the parameters were not always normally distributed (decision noise:
364 vehicle: $p= 0.85$, APO: $p<0.0001$; vehicle: $p<0.0001$, FLU: $p=0.09$), therefore we will report
365 both parametric and non-parametric statistics results for RL parameters. Increasing dopamine
366 receptor activity with apomorphine administration increased inverse temperature and decreased
367 random noise (**Figure 5C**, Wilcoxon matched-pairs test, $p= 0.0165$; paired t-test, $p= 0.053$).
368 Decreasing dopamine receptor activity with flupenthixol administration decreased inverse
369 temperature and increased random noise (**Figure 5D**, Wilcoxon matched-pairs test, $p= 0.0253$;
370 paired t-test, $p= 0.053$). Neither dopamine agonist (apomorphine) nor dopamine antagonist
371 (flupenthixol) significantly influenced learning rate or choice kernel compared to vehicle
372 controls.
373
374 In norepinephrine modulation conditions, the RLCK model (learning + choice kernel,) and the
375 RLCKγ (asymmetrical learning and bias) model also best fit the behavioral data (**Figure 5F, G,**
376 average AIC across NE conditions: model 1 random: 66898.83; model 2 WSLS: 56061.3; model

3 RL: 53194.59; model 4 RLCK: 50035.02; model 5 RLγ: 52664.48; model 6 RLCKγ:50027.82). However, since the RLCKγ model did not significantly improve the model fitting, we decided to use the more parsimonious RLCK model for the NE condition as well. We found that modulating beta-noradrenergic receptor activity also resulted in changes in the reinforcement learning model parameter. Up-regulating norepinephrine activity with isoproterenol significantly decreased the learning rate (**Figure 5H**, Wilcoxon matched-pairs test, p= 0.0001; paired t-test, p= 0.0008). This result is consistent with the decreased win-stay and decreased lose-shift behaviors when administered isoproterenol because the lower learning rate under isoproterenol resulted in reduced outcome sensitivity (**Supplementary figure 1C**), which means learning less from either a rewarded or a non-rewarded outcome. There were no significant changes in other parameters when administered isoproterenol compared to vehicle control. We did not find any significant changes in any reinforcement learning parameters when administered propranolol compared to vehicle control. Together, these results suggested distinct roles of dopaminergic drugs in modulating exploration via decision noise and isoproterenol in modulating exploration via outcome sensitivity.

Discussion

In this study, we pharmacologically up- and down-regulated dopamine and beta-noradrenergic receptor activity and examined the modulatory effects on exploration in an explore/exploit task. We used a combination of computational modeling approaches to characterize changes in exploration and latent cognitive variables that could contribute to exploration. Modulating dopamine receptor activity revealed a bi-directional modulatory effect on exploration - increasing dopamine receptor activity decreased the level of exploration and decreasing dopamine receptor activity increased the level of exploration. Modulation of beta-noradrenergic receptor activity also influenced how much animals explored - beta-noradrenergic receptor antagonist (propranolol) decreased the level of exploration. However, when examining the modulatory effect of beta-noradrenergic receptor activity on exploration across sexes, we found that modulatory effect was sex-different: isoproterenol selectively decreased exploration in females whereas propranolol decreased exploration in males. Previously, the role of dopamine and norepinephrine in mediating exploration were often examined in isolation with different task designs. This study allowed the direct comparison of the modulatory effect of dopamine and norepinephrine on the transition between exploration and exploitation in the same task across the same animals.

Dopamine's function as a neuromodulator has been described in two key ways. First, the phasic activation of midbrain dopamine neurons is thought to be a key contributor to reinforcement learning [30,59–62]. Numerous studies across species have shown that dopamine neurons encoded the reward prediction error (RPE) [30,31]. Ventral tegmental dopamine neuronal activity and release in ventral striatum is proposed to reflect this error, which is used to update action values [30–32]. Consistent with this view, some previous pharmacological investigations in human decision making have observed changes in model-derived learning rates that directly correlate with dopamine synthesis agonism via L-DOPA [63,64] and dopamine receptor D2/D3 antagonism via amisulpride [26]. However, this framework does not necessarily account for a large body of literature describing a second way to view dopamine's neuromodulatory function, as tonic levels influencing motivation, vigor, and cognitive flexibility [48,65], especially when pharmacologically targeted in frontal cortex. Because pharmacological agents are by definition

modulating receptor activity over long periods of time, this role for tonic dopamine is poised to be computationally described. Indeed, when pharmacologically or genetically modulating tonic dopamine function, many studies have noted profound changes in the decision “noise” (inverse temperature) or preservative errors - consistent with a role for dopamine modulation in cognitive flexibility in earlier work [35,38,66–69]. Two prior examples combining pharmacological manipulation and computational modeling in animal tasks explicitly testing the explore-exploit tradeoff particularly highlight this. Cinotti et al. found that systemic dopamine blockade with flupenxithol increased the amount of random exploratory choices without affecting learning rate in a rat three-arm bandit task [38], particularly in situations where uncertainty was manipulated. Similarly, Ebitz et al. reported that systemic cocaine, which blocks dopamine reuptake, reduced flexibility and regulated tonic exploration in rhesus macaques [69]. Likewise, we find that systemic activation of dopamine receptors significantly decreased exploration as defined by our Hidden Markov model and decision noise as described by our reinforcement learning model, while antagonism of these receptors increased these computational parameters measuring flexible behavior. Our data thus provide additional support for the idea that dopamine modulation across the brain provides more than a learning signal, but also a key signal tuning the precision or flexibility of behavior.

As noted in our introduction, the roles of dopamine and norepinephrine are sometimes conflated in the computational neuroscience literature. This is in part because modulation of these systems can be correlated, and noninvasive measures of catecholamine function, for example blink rate and pupil diameter, could be ascribed to the endogenous activity of either neuromodulator (although pharmacological action may target one or the other) [26,53,70–72]. This is further complicated by the fact that some studies examining the relationship between dopamine and exploration have used pharmacological manipulations which simultaneously influence both tonic dopamine and norepinephrine levels, such as psychostimulants or L-DOPA [63,69,73–75]. This makes it challenging to dissociate the modulatory roles of dopamine and norepinephrine in exploration i.e. how dopamine and norepinephrine distinctively contribute to changes in reinforcement learning model parameters, and addressing this challenge was a major goal of our study.

Surprisingly, we found that when modulating beta-noradrenergic receptor activity changed the level of exploration, it did so via changes in outcome sensitivity. Previous studies have proposed the locus coeruleus-norepinephrine (LC-NE) system to regulate exploration by changing decision noise. Kane et al. chemogenetically stimulated locus coeruleus activity in rats and found that increased LC tonic activity increased in increased exploration, best explained by an increase in decision noise [9,76]. Other studies have echoed the findings that norepinephrine levels predicted the level of random exploration, or decision noise, in the system [53,71,77]. However, in this study we did not find any changes in the noise parameter (inverse temperature, or how closely you adhere to your value rule) associated with up- and down-regulation of beta-noradrenergic activity. Instead, the results revealed some learning rate changes when increasing beta-noradrenergic receptor activity. Model-free analysis on outcome sensitivity suggested that increasing or decreasing beta-noradrenergic receptor activity bi-directionally changed outcome sensitivity. Propranolol increased outcome sensitivity and decreased exploration, which means the animals were more likely to switch choice after loss and more likely to repeat rewarded choices. In contrast isoproterenol decreased outcome sensitivity, as seen in reduced win-stay and lose-shift behavior. This is consistent with the decreased learning rate when administered isoproterenol because a decrease in learning rate is associated with decreased value learning from each outcome. Where might this strange association between noradrenergic modulation, exploration, and outcome sensitivity emerge from? We have previously shown that exploration as defined by our Hidden Markov model can be driven by either changes in learning rate or changes in inverse temperature [4], supporting the idea that the exploration we see here could be driven by long-term changes in the influence of outcomes. Work focused on beta noradrenergic signaling in the context of aversive learning has highlighted a role for noradrenergic signaling in extinction [78], over a phasic timeframe of minutes to hours. This suggests the intriguing possibility that the effect of noradrenergic blockade here may be to prevent extinction of learned values. In short, although the most common associations in the computational neuroscience literature are between dopamine and reward learning versus norepinephrine and decision noise, our current data suggests a more nuanced role for each of these modulators.

Adding an additional layer of nuance to these findings, we find that sex differences are a major axis of variability in the effect of beta-noradrenergic receptor modulation on the explore-exploit

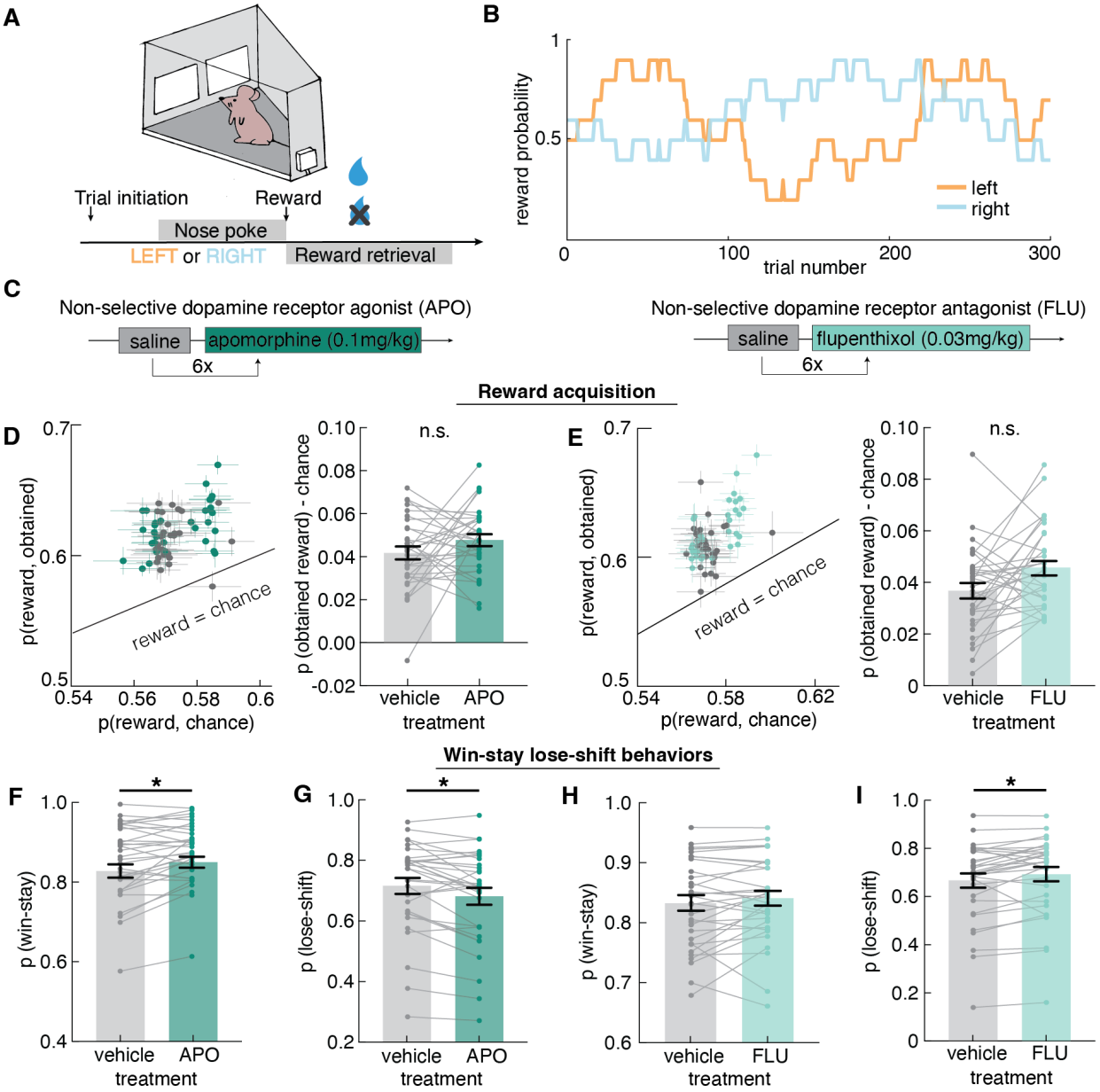
tradeoff. Increasing NE receptor activity with the beta agonist isoproterenol decreased exploration only in males, while decreasing NE receptor activity with the beta antagonist propranolol decreased exploration especially in females. One possible explanation is that the relationship between exploration and NE receptor activity was nonlinear but instead, followed an inverted U-shape. Such inverted U-shape correlation has been suggested between cognitive performance and NE level [23,27,79]. Since females have been reported to have higher tonic NE levels due to both function and structural differences [50,80], it is possible that the baseline difference in tonic NE level across sexes could posit individuals on different parts of the inverted U-curve, where increasing or decreasing NE could affect exploration in one sex without impacting the other.

Taken together, this study demonstrates the distinct role of dopamine and norepinephrine in tuning the exploration-exploitation tradeoff when learning about an uncertain environment. Dopamine modulated exploration via decision noise and norepinephrine modulated exploration via outcome sensitivity. The systemic manipulation of dopamine and norepinephrine receptor activity did not allow for more precise modulation of phasic changes in catecholamine levels. However, the computational tool that we presented here can complement higher resolution neural recording techniques to examine for phasic changes in dopamine and norepinephrine activity on a trial-by-trial basis. Finally, this study also highlighted that the modulatory effects of catecholamine on exploration-exploitation tradeoff are strongly influenced by sex and sex-linked mechanisms. Therefore, future research should take into account sex-correlated individual variability when examining the neural correlates of exploration.

509 **Figures and figure legends**

510

511 **Figure 1**



512

513

514 **Figure 1. Modulation of dopamine receptor activity affected reward-directed behaviors.**
515 **Up-regulating dopamine activity increased stickiness in choice behaviors regardless of**
516 **outcomes.**

A) Schematic of the mouse touchscreen chamber and the trial structure of the two-armed spatial restless bandit task. **B)** An example of the dynamic reward contingency showing the changing reward probabilities associated with each option. **C)** The dopaminergic drug administration schedule and drug dosage used to modulate dopamine receptor activity. A non-selective dopamine receptor agonist apomorphine (0.1mg/kg) and a non-selective dopamine receptor antagonist flupenthixol (0.03mg/kg) were systemically administered to respectively up- and down-regulate dopamine receptor activity. 0.9% saline was used as the vehicle control. Control and drug sessions were interleaved and repeated for six sessions each. **D)** (left) Average probability of obtaining reward compared to the chance level probability of reward for each animal under apomorphine (APO) and vehicle (dot). Error bars depict mean ± SEM across sessions for each animal. (right) probability of obtaining reward over chance level across APO (dark green) and vehicle (gray). **E)** (left) Average probability of obtaining reward compared to the chance level probability of reward for each animal under flupenthixol (FLU) and vehicle (dot). Error bars depict mean ± SEM across sessions for each animal. (right) probability of obtaining reward over chance level across FLU (light green) and vehicle (gray). **F)** Average probability of win-stay on vehicle control or APO. **G**) Average lose-shift on vehicle control or APO. **H)** Average probability of win-stay on vehicle control or FLU. **I**) Average lose-shift on vehicle control or FLU. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals unless specified otherwise.

Figure 2

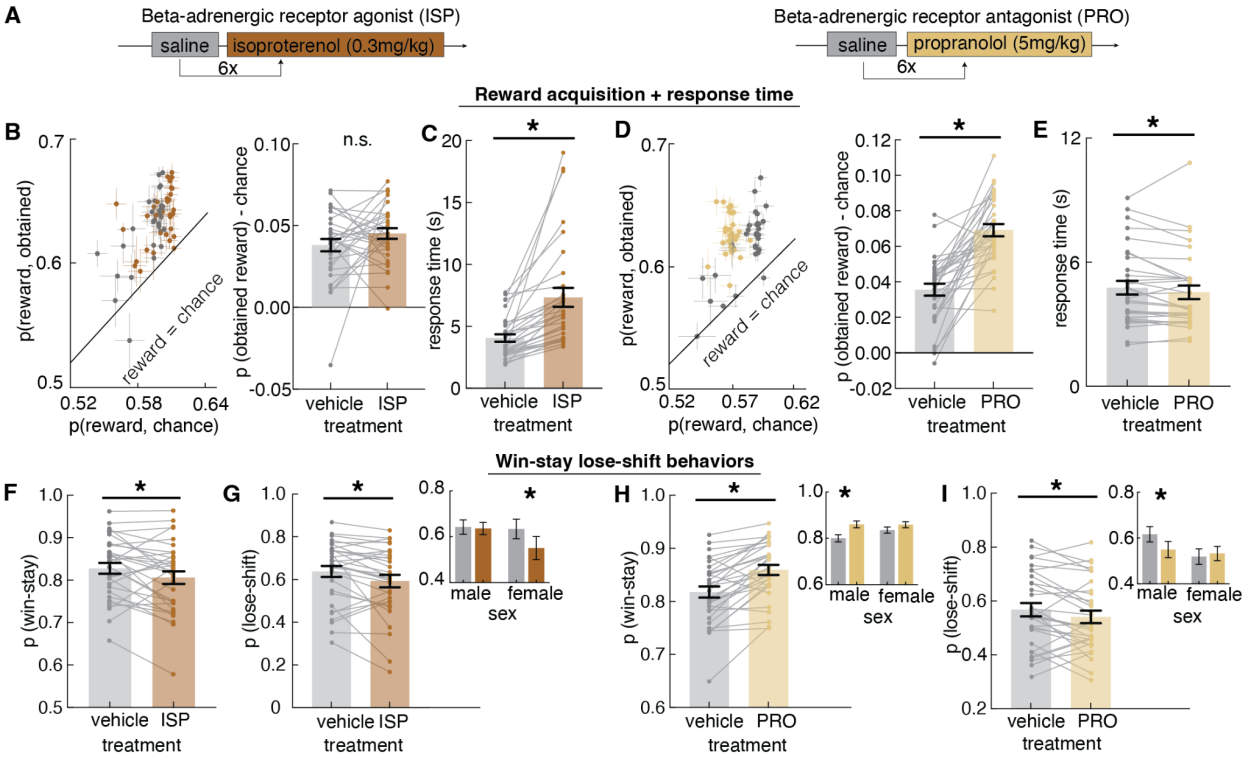


Figure 2. The modulatory effects of beta-noradrenergic receptor activity on reward-directed behavior was also influenced by sex.

A) The noradrenergic drug administration schedule and drug dosage used to modulate noradrenergic receptor activity. A beta-noradrenergic receptor agonist isoproterenol (0.3mg/kg) and a beta-noradrenergic receptor antagonist propranolol (5mg/kg) were systemically administered to respectively up- and down-regulate norepinephrine activity. 0.9% saline was used as the vehicle control. Control and drug sessions were interleaved and repeated for six sessions each. **B)** (left) Average probability of obtaining reward compared to the chance level probability of reward for each animal under isoproterenol (ISP) and vehicle (dot). Error bars depict mean ± SEM across sessions for each animal. (right) probability of obtaining reward over chance level across ISP (dark brown) and vehicle (gray). **C)** Average response time under ISP and vehicle. ISP significantly increased response time. **D)** (left) Average probability of obtaining reward compared to the chance level probability of reward for each animal under propranolol (PRO) and vehicle (dot). Error bars depict mean ± SEM across sessions for each animal. (right) probability of obtaining reward over chance level across PRO (light brown) and vehicle (gray).

560 **E)** Average response time under PRO and vehicle. PRO significantly decreased response time.
561 **F)** Average probability of win-stay under ISP and vehicle. **G)** Average probability of lose-shift
562 under ISP and vehicle. (inset) probability of lose-shift across treatments across sexes. Decrease
563 in lose-shift under ISP was primarily driven by changes in females (interaction term). **H)**
564 Average probability of win-stay under PRO and vehicle. (inset) probability of win-stay across
565 treatments across sexes. Increase in win-stay under PRO was primarily driven by changes in
566 males (interaction term). **I)** Average probability of lose-shift under PRO and vehicle. (inset)
567 probability of lose-shift across treatments across sexes. Decrease in lose-shift under PRO was
568 primarily driven by changes in males (interaction term). * indicates $p < 0.05$. Graphs depict mean
569 ± SEM across animals unless specified otherwise.

Figure 3

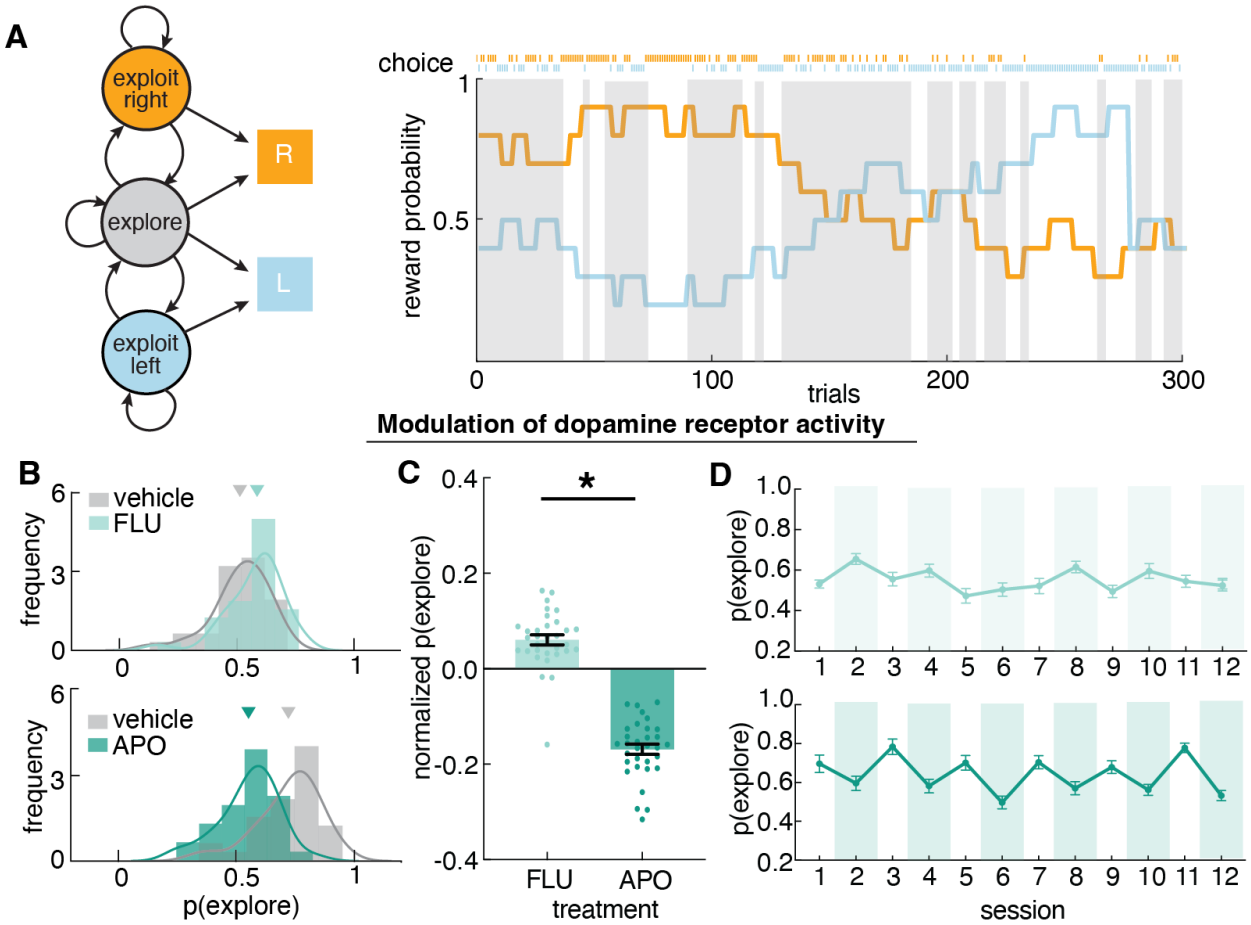


Figure 3. Up- and down-regulation of dopamine activity had bidirectional effects on the level of exploration.

A) (left) Structure of a Hidden Markov model (HMM) that modeled exploration and exploitation as latent goal states undering observed choices. This model incorporates two exploit states for two choices and one explore state where choices were uniformly distributed across options. (right) Reward probabilities, choices, and HMM labels for an example session of 300 trials for a given mouse. Shaded areas demonstrated HMM-labeled exploratory choices. **B)** Distribution of the percentage of HMM-labeled exploratory choices under flupenthixol (FLU)/ vehicle (top) and apomorphine (APO)/vehicle (bottom). **C)** Probability of exploration for dopamine antagonist (FLU) and agonist (APO), normalized by their vehicle control. Decreasing dopamine activity increased exploration and increasing dopamine activity decreased exploration. **D)** Probability of exploration by session with vehicle and drug session interleaved (flupenthixol: top;

585 apomorphine: bottom). Drug administration sessions are in colored shades. * indicates $p < 0.05$.
586 Graphs depict mean ± SEM across animals.
587

Figure 4

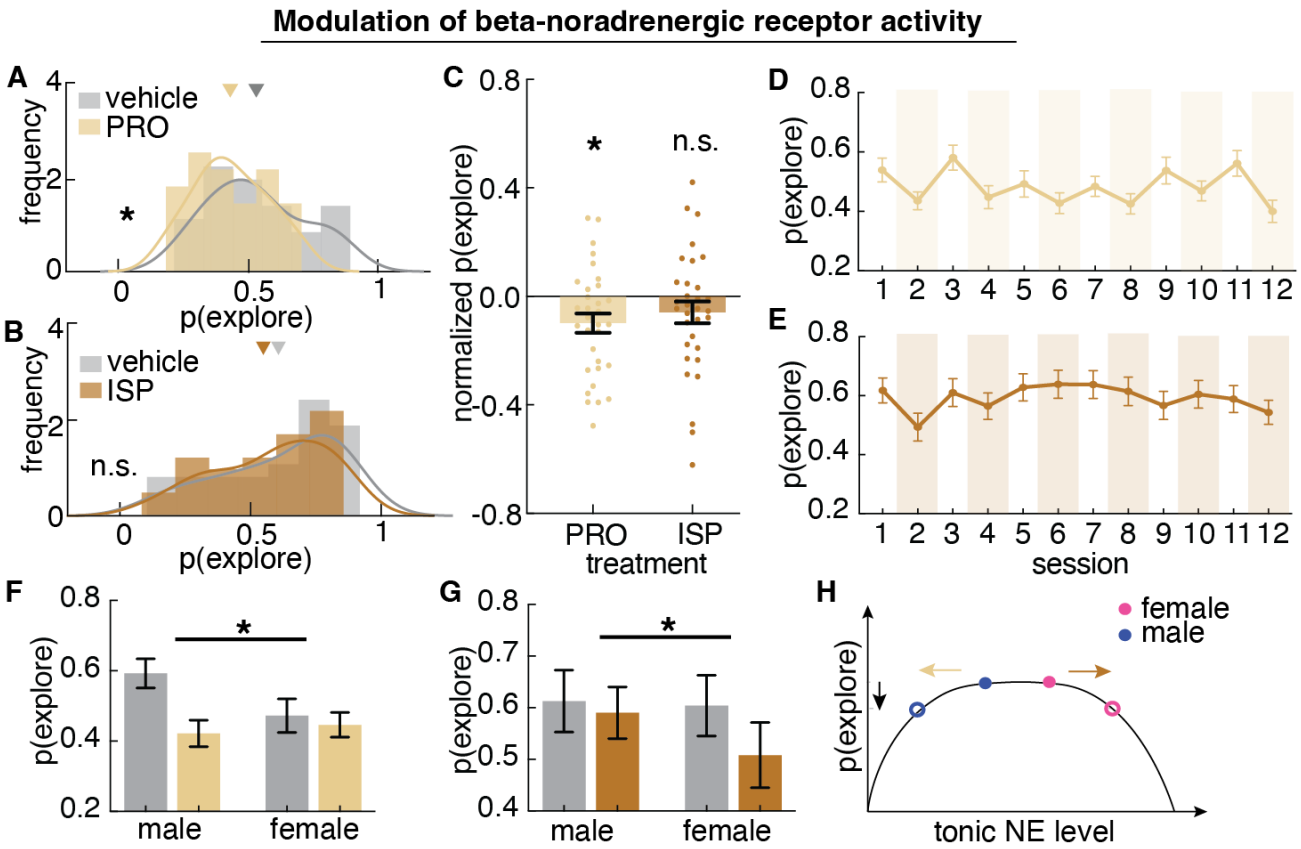


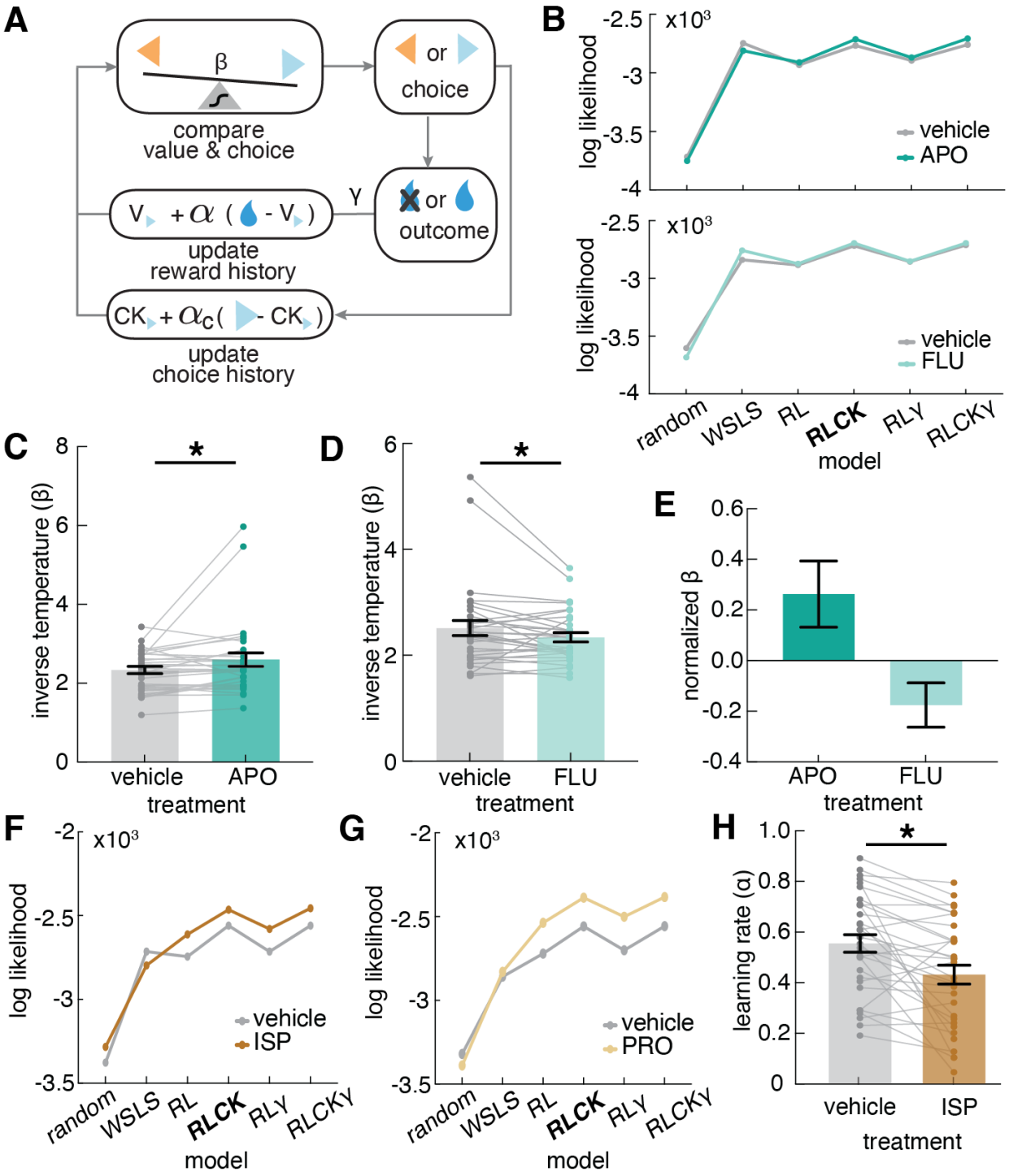
Figure 4. Downregulating norepinephrine activity influenced exploration but the effect was modulated by sex.

A) Distribution of the percentage of HMM-labeled exploratory choices under propranolol (PRO)/ vehicle. **B)** Distribution of the percentage of HMM-labeled exploratory choices under isoproterenol (ISP)/vehicle. **C)** Probability of exploration for beta-noradrenergic antagonist (PRO) and agonist (ISP), normalized by their vehicle control. Propranolol decreased the level of exploration. **D)** Probability of exploration by session with vehicle and PRO session interleaved. Drug administration sessions are in colored shades. **E)** Probability of exploration by session with vehicle and ISP session interleaved. Drug administration sessions are in colored shades. **F)** Probability of exploration across ISP and vehicle across sexes. ISP significantly decreased the level of exploration in males **G)** Probability of exploration across PRO and vehicle across sexes. The effect of PRO on exploration was primarily driven by a significant decrease in exploration under PRO in females. **H)** Cartoon of proposed inverted U-shape relationship between tonic

604 norepinephrine (NE) level and level of exploration. Sex difference in baseline NE level may
605 contribute to differential modulation of NE agonist and antagonist across sexes. * indicates $p <$
606 0.05. Graphs depict mean ± SEM across animals.

607 **Figure 5**

608



609

610 **Figure 5. Modulation of dopamine and noradrenergic receptor activity led to changes in**

611 **different reinforcement learning (RL) model parameters.**

A) A diagram of reinforcement learning (RL) parameters that capture learning rate (α), asymmetric learning (γ),choice bias (α_c), inverse temperature/decision noise (β). The RL models tested included one or more of the above parameters. **B)** Model agreement for six RL models using log likelihood for vehicle and apomorphine (APO) condition (top), and vehicle and flupenthixol condition (bottom). The three-parameter RLCK model was the best fit and most parsimonious model for behaviors. **C)** Average inverse temperature (β) across apomorphine and vehicle. **D)** Average inverse temperature (β) across flupenthixol and vehicle. **E**) Agonizing (apomorphine) and antagonizing (flupenthixol) dopamine activity revealed bidirectional effect on inverse temperature. **F**) Model agreement for six RL models using log likelihood for vehicle and isoproterenol (ISP) condition **G**) Model agreement for six RL models using log likelihood for vehicle and propranolol (PRO) condition. **H**) Average learning rate (α) across isoproterenol and vehicle. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

Supplementary figure 1

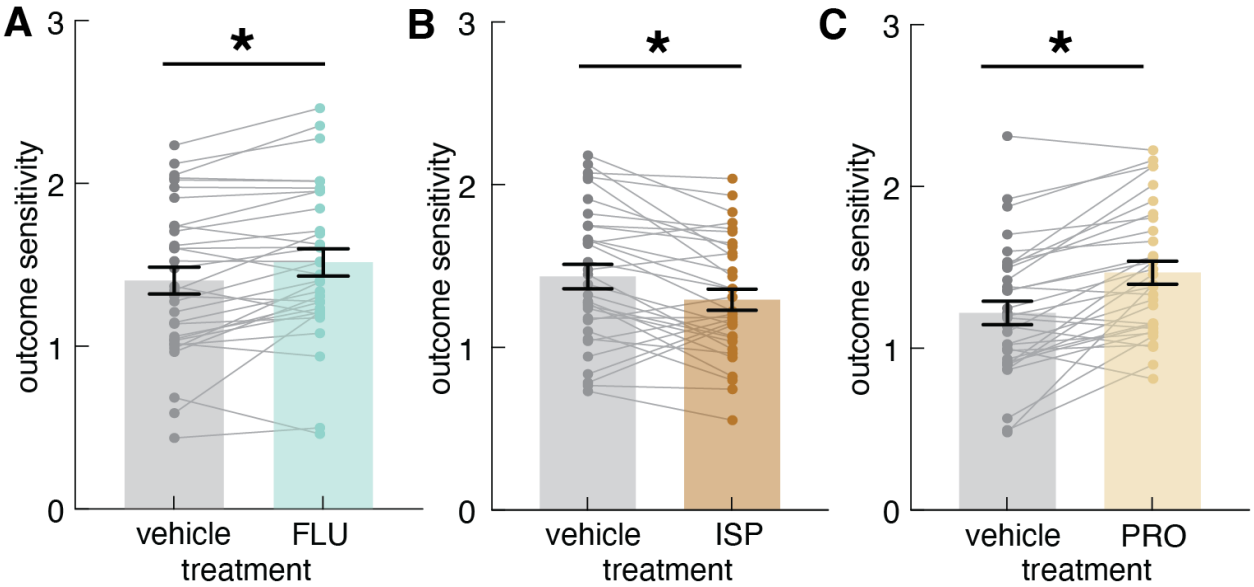


Figure S1. Effect of dopamine and norepinephrine modulation on outcome sensitivity. Related to Figure 1.

A) Average outcome sensitivity across flupenxithol (FLU) and vehicle. **B**) Average outcome sensitivity across isoproterenol (ISP) and vehicle. **C**) Average outcome sensitivity across propranolol (PRO) and vehicle. * indicates $p < 0.05$. Graphs depict mean ± SEM across animals.

Supplementary figure 2

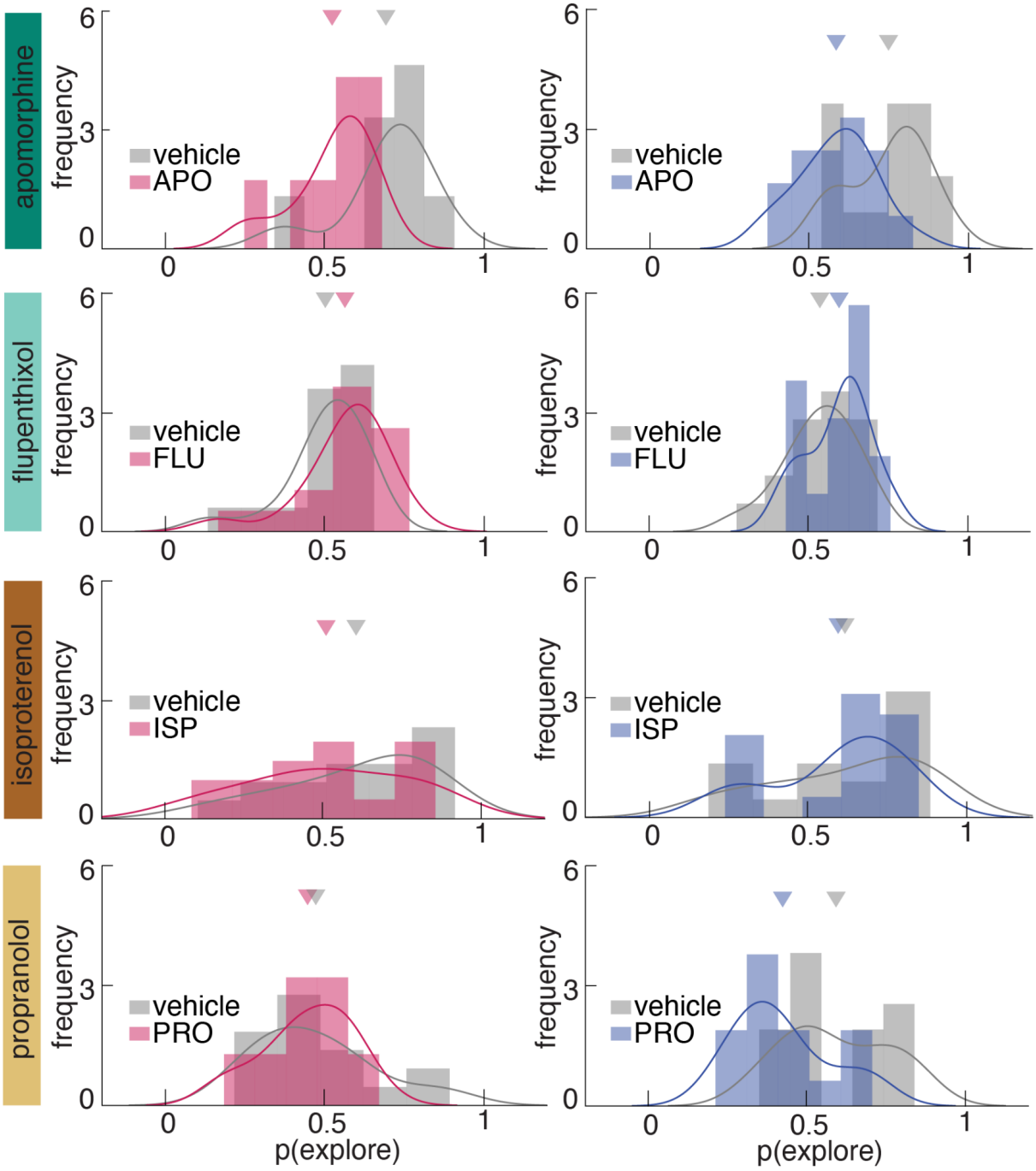


Figure S2. Effect of dopamine and norepinephrine modulation on exploration in female and male mice. Related to Figure 3,4.

641 **Methods**

642 Animals. Thirty-two BL6129SF1/J mice (16 males and 16 females) were obtained from Jackson
643 Laboratories (stock #101043). Mice arrived at the lab at 7 weeks of age and adapted to 0900-
644 2100 hours reversed light cycle throughout the testing phase. Mice were housed in groups of four
645 with ad libitum access to water while being mildly food restricted (85-95% of free feeding
646 weight) before the start of the experiment (12 weeks) and during the experiment. All animals
647 were cared for according to the guidelines of the National Institution of Health and experimental
648 protocols were as approved by the Institutional Animal Care and Use Committee (IACUC) of the
649 University of Minnesota.

650

651 Apparatus. Sixteen identical triangular touchscreen operant chambers (Lafayette Instrument Co.,
652 Lafayette, IN) were used for training and testing. Two walls were black acrylic plastic. The third
653 wall housed the touchscreen and was positioned directly opposite the magazine. The magazine
654 provided liquid reinforcers (50% Ensure) delivered by a peristaltic pump, typically 7ul (280 ms
655 pump duration). ABET-II software (Lafayette Instrument Co., Lafayette, IN) was used to
656 program operant schedules and to analyze all data from training and testing.

657

658 Behavioral task. *Two-armed spatial restless bandit task.* Animals were trained to perform a two-
659 armed spatial restless bandit task in the touchscreen operant chamber. Each trial, animals were
660 presented with two identical squares on the left and right side of the screen. Nose poke to one of
661 the target locations on the touchscreen was required to register a response. Each location is
662 associated with some probability of reward, which changes independently over time. For every
663 trial, there is a 10% chance that the reward probability of a given arm will increase or decrease
664 by 10%. All the walks were generated randomly with a few criteria: 1) the overall reward
665 probabilities of two arms are within 20% of each other, preventing one arm being overly better
666 than the other, 2) the reward probability cannot go down to 0% or go up to 100%, 3) there are no
667 30 consecutive trials where the reward probabilities of both arms are lower than 20% to ensure
668 motivation.

669

670 Animals ran a simple deterministic schedule on Monday to re-adapt to the operant chamber after
671 weekends off and ran a different restless bandit walk 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. On average across all sessions, animals performed 276.5 trials with a standard deviation of 8.6 trials (male average: 253.7 trials, std = 15.4; female average 299.3 trials, std = 0.74. Data was recorded by the ABET II system and was exported for further analysis. All computational modeling was conducted using python.

Drug administration. To assess the effect of norepinephrine and dopamine receptor activity on explore behaviors, we used the following four different drugs to increase or decrease beta-noradrenergic or dopaminergic receptor activity systemically: a beta-noradrenergic receptor agonist isoproterenol (0.3mg/kg), a beta-noradrenergic receptor antagonist propranolol (5mg/kg), a non-selective dopamine receptor agonist apomorphine (0.1mg/kg), and a non-selective dopamine receptor antagonist flupenthixol (0.03mg/kg).

All drugs were fully dissolved in 0.9% saline and protected from light. Animals received intraperitoneal (IP) injection of drug or saline (control) at an injection volume of 5ml/kg right before the experiment with the exception of apomorphine being administered 20 minutes before the experiment due to the immediate drug influence of locomotor behaviors. Drug and saline were administered in alternating sessions using a within-subject so that every animal received each of the four drugs. The dopamine receptor agonist and antagonist were administered counterbalanced across animals. Half of the animals (n = 8) were randomly selected to receive apomorphine first, the other received flupenthixol first. Dosage chosen was based on previous studies on the effect of drugs on cognitive functions [38–44]. Doses for each drug were chosen from previous studies as the lowest doses necessary to produce alterations in cognitive functions including decision making, learning and exploration with minimal influence on locomotion.

Data Analysis

General analysis techniques. Data was analyzed with custom PYTHON, and Prism 8 scripts. Generalized linear mixed models (GLMMs), ANOVA, and t-test were used to determine the fixed effects of drug, sex, and interaction term, accounting for random effects of animal identity and session, unless otherwise specified. P values were compared against the standard $\alpha = 0.05$ threshold. The sample size is n = 32 (16 males and 16 females) for all statistical tests. No animal

was excluded from the experiment. All statistical tests used and statistical details were reported in the results. All figures depict mean ± SEM.

Generalized Linear Mixed Models (GLMMs). In order to determine whether drug and sex predicted the reward acquisition performance, response time, probability of win-stay, probability of lose-shift, outcome sensitivity, probability of exploration, we fitted a series of generalized linear mixed model to examine the fixed effects of drug, sex, and interaction term, with animal identity and session number as random effects.

$$Y = \beta_0 + \beta_1(drug) + \beta_2(sex) + \beta_3(drug)(sex) + u_1(animal\ ID) + u_2(session) \quad \textbf{[1]}$$

Where Y is the dependent variable (obtained reward, response time, win-stay, lose-shift, outcome sensitivity, exploration). In this model, β1 describes the main effect of drug and β2 describes the main effect of sex. β3 captures any interaction effect between drug and sex. u1 and u2 capture the random effect of animal identity and session respectively.

Outcome Sensitivity. In order to examine how much of the switching behavior was sensitive to reward outcome, we calculated the difference between switching given no reward and switching given reward control by the total amount of switching.

$$outcome\ sensitivity = \frac{p(switch|no\ reward) - p(switch|reward)}{p(switch)} \quad \textbf{[2]}$$

Hidden Markov model (HMM). In order to identify when animals were exploring, we fit a Hidden Markov model (HMM) to the animals' choice sequence. Our model consisted of two hidden states - explore state and exploit state, which are defined by the probability of making each choice (k, out of N_k possible options) and the probability of transitioning from one state to another. The HMM models exploration as a uniform distribution over choices, i.e. the emissions model for the explore state was uniform across options. This is the maximum entropy distribution of a categorical variable, which makes the few assumptions of true distribution of choices during exploration and therefore does not bias the model towards or away from any particular type of high-entropy choice period. In the exploitation state, subjects repeatedly sample the same choice, therefore the exploit states only permit one choice, i.e. exploit-left state only emits left choices and exploit-right state only emits right choice.

The latent states in this model are Markovian, meaning that they are time-independent. They depend only on the most recent state. the HMM model estimates a transition matrix to fit the behavior by mapping of past and future states. This matrix is a system of stochastic equations describing the 1-time-step probability of transitioning between every combination of states. In our model, there were 3 possible states (2 exploit states and 1 explore state). The 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 also 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 [81]. 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, the initial distribution started with p (explore state) = 1. The algorithm was reinitialized with random seeds 10 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 posteriori sequence of latent states.

Reinforcement learning (RL) models. We fitted five reinforcement learning (RL) models that could potentially characterize animals' choice behaviors, with the structure of each RL model detailed below. AIC weights were calculated from AIC values across all treatment groups and compared across models to determine the best model with the highest relative likelihood.

The first model (random) 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:

$$p_t^L = b \qquad [3] \quad \textit{"random"}$$

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:

$$p_t^k = \begin{cases} 1 - \frac{\epsilon}{2}, if\ (c_{t-1} = k\ and\ r_{t-1} = 1\ OR\ c_{t-1} \neq k\ and\ r_{t-1} = 0) \\ \frac{\epsilon}{2}, if\ (c_{t-1} \neq k\ and\ r_{t-1} = 1\ OR\ c_{t-1} = k\ and\ r_{t-1} = 0) \end{cases} \quad [4]\ \textit{“noisy WSLS”}$$

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 (RL) 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:

$$Q_{t+1}^k = Q_t^k + \alpha(r_t - Q_t^k) \quad [5]\ \textit{”RL”}$$

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 (*RLCK*) 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:

$$CK_{t+1}^k = CK_t^k + \alpha_c(a_t^k - CK_t^k) \quad [6]\ \textit{“RLCK”}$$

, 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*:

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

, where β is the inverse temperature, capturing the decision noise outside value and choice bias.

The fifth model (*RLγ)* is an asymmetrical learning model that incorporates an asymmetric learning scalar parameter that scales the learning rate on the trials where there is no reward. The choice selection is also based on a Softmax probability distribution where the inverse temperature determines the decision noise in the system.

$$Q_{t+1}^k = \begin{cases} Q_t^k + \alpha\left(r_t - Q_t^k\right), r_t = 1 \\ Q_t^k + \gamma \times \alpha\left(r_t - Q_t^k\right), r_t = 0 \end{cases} \qquad [7]\ \text{“}RL\gamma\text{”}$$

The sixth model (*RLCKγ*) is the combination of the fourth and the fifth model, incorporating both asymmetrical learning and choice kernel. This model included a learning rate, an asymmetric learning scalar, a choice kernel, and an inverse temperature parameter.

Competing Interests

The authors have declared no competing interests.

Acknowledgements

Thank you to Dr. David Redish and Madison J Merfeld for comments improving this manuscript. Thank you to Anila Bano for helping with the experiment.

Author contributions

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

Funding

This work was supported by NIMH R01 MH123661 (N.M.G), NIMH P50 MH119569 (N.M.G), and NIMH T32 training grant MH115886 (C.S.C), 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).

824 **References**

825 1. Grissom NM, Reyes TM. Let's call the whole thing off: evaluating gender and sex
826 differences in executive function. Neuropsychopharmacology. 2019. 2019.
827 2. Malloy-Diniz LF, Miranda DM, Grassi-Oliveira R. Editorial: Executive Functions in
828 Psychiatric Disorders. Front Psychol. 2017;8:1461.
829 3. Elliott R. Executive functions and their disorders. Br Med Bull. 2003;65:49–59.
830 4. Chen CS, Knep E, Han A, Ebitz RB, Grissom N. Sex differences in learning from
831 exploration. Elife. 2021;10.
832 5. Chen CS, Ebitz RB, Bindas SR, Redish AD, Hayden BY, Grissom NM. Divergent
833 Strategies for Learning in Males and Females. Curr Biol. 2021;31:39–50.e4.
834 6. Ebitz RB, Albarran E, Moore T. Exploration Disrupts Choice-Predictive Signals and Alters
835 Dynamics in Prefrontal Cortex. Neuron. 2018;97:475.
836 7. Rothenhoefer KM, Costa VD, Bartolo R, Vicario-Feliciano R, Murray EA, Averbeck BB.
837 Effects of Ventral Striatum Lesions on Stimulus-Based versus Action-Based Reinforcement
838 Learning. J Neurosci. 2017;37:6902–6914.
839 8. Sutton RS, Barto AG. Reinformcent Learning: An Introduction. 1998.
840 9. Wilson RC, Bonawitz E, Costa VD, Ebitz RB. Balancing exploration and exploitation with
841 information and randomization. Curr Opin Behav Sci. 2021;38:49–56.
842 10. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory
843 decisions in humans. Nature. 2006;441:876–879.
844 11. Niv Y, Joel D, Meilijson I, Ruppin E. Evolution of reinforcement learning in uncertain
845 environments: A simple explanation for complex foraging behaviors. Adapt Behav.
846 2002;10:5–24.
847 12. Jepma M, Schaaf JV, Visser I, Huizenga HM. Uncertainty-driven regulation of learning and
848 exploration in adolescents: A computational account. PLoS Comput Biol.
849 2020;16:e1008276.
850 13. Gershman SJ. Uncertainty and Exploration. Decision (Wash D C). 2019;6:277–286.
851 14. Kim H, Lee D, Shin Y-M, Chey J. Impaired strategic decision making in schizophrenia.
852 Brain Res. 2007;1180:90–100.
853 15. Mussey JL, Travers BG, Klinger LG, Klinger MR. Decision-making skills in ASD:
854 performance on the Iowa Gambling Task. Autism Res. 2015;8:105–114.
855 16. Addicott MA, Pearson JM, Schechter JC, Sapyta JJ, Weiss MD, Kollins SH. Attention-
856 deficit/hyperactivity disorder and the explore/exploit trade-off. Neuropsychopharmacology.
857 2020. 11 October 2020. https://doi.org/10.1038/s41386-020-00881-8.
858 17. Morris LS, Baek K, Kundu P, Harrison NA, Frank MJ, Voon V. Biases in the Explore-
859 Exploit Tradeoff in Addictions: The Role of Avoidance of Uncertainty.
860 Neuropsychopharmacology. 2016;41:940–948.
861 18. Kaske EA, Chen CS, Meyer C, Yang F, Ebitz B, Grissom N, et al. Prolonged physiological
862 stress is associated with a lower rate of exploratory learning that is compounded by
863 depression. Biol Psychiatry Cogn Neurosci Neuroimaging. 2022. December 2022.
864 https://doi.org/10.1016/j.bpsc.2022.12.004.
865 19. Kobayashi K. Role of catecholamine signaling in brain and nervous system functions: new
866 insights from mouse molecular genetic study. J Investig Dermatol Symp Proc. 2001;6:115–
867 121.
868 20. Money KM, Stanwood GD. Developmental origins of brain disorders: roles for dopamine.

Front Cell Neurosci. 2013;7:260.
21. Ressler KJ, Nemeroff CB. Role of norepinephrine in the pathophysiology of neuropsychiatric disorders. CNS Spectr. 2001;6:663–666, 670.
22. Williams OOF, Coppolino M, George SR, Perreault ML. Sex Differences in Dopamine Receptors and Relevance to Neuropsychiatric Disorders. Brain Sci. 2021;11.
23. Aston-Jones G, Cohen JD. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci. 2005;28:403–450.
24. Bouret S, Sara SJ. Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. Trends Neurosci. 2005;28:574–582.
25. Cohen JD, McClure SM, Yu AJ. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007;362:933–942.
26. Cremer A, Kalbe F, Müller JC, Wiedemann K, Schwabe L. Disentangling the roles of dopamine and noradrenaline in the exploration-exploitation tradeoff during human decision-making. Neuropsychopharmacology. 2022. 15 December 2022. https://doi.org/10.1038/s41386-022-01517-9.
27. Redish AD, Jensen S, Johnson A, Kurth-Nelson Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychol Rev. 2007;114:784–805.
28. Seamans JK, Yang CR. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. Prog Neurobiol. 2004;74:1–58.
29. Ranjbar-Slamloo Y, Fazlali Z. Dopamine and Noradrenaline in the Brain; Overlapping or Dissociate Functions? Front Mol Neurosci. 2019;12:334.
30. Schultz W. Predictive reward signal of dopamine neurons. J Neurophysiol. 1998;80:1–27.
31. Niv Y. Reinforcement learning in the brain. J Math Psychol. 2009;53:139–154.
32. Schultz W. Updating dopamine reward signals. Curr Opin Neurobiol. 2013;23:229–238.
33. Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005;46:681–692.
34. Amemiya S, Kubota N, Umeyama N, Nishijima T, Kita I. Noradrenergic signaling in the medial prefrontal cortex and amygdala differentially regulates vicarious trial-and-error in a spatial decision-making task. Behav Brain Res. 2016;297:104–111.
35. Humphries MD, Khamassi M, Gurney K. Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. Front Neurosci. 2012;6:9.
36. Warren CM, Wilson RC, van der Wee NJ, Giltay EJ, van Noorden MS, Cohen JD, et al. The effect of atomoxetine on random and directed exploration in humans. PLoS One. 2017;12:e0176034.
37. Grossman CD, Bari BA, Cohen JY. Serotonin neurons modulate learning rate through uncertainty. bioRxiv. 2020:2020.10.24.353508.
38. Cinotti F, Fresno V, Aklil N, Coutureau E, Girard B, Marchand AR, et al. Dopamine blockade impairs the exploration-exploitation trade-off in rats. Sci Rep. 2019;9:6770.
39. Grigoryan GA. Serotonin and Impulsivity (animal experiments). Neurosci Behav Physiol. 2012;42:885–894.
40. Ichihara K, Nabeshima T, Kameyama T. Opposite effects induced by low and high doses of apomorphine on single-trial passive avoidance learning in mice. Pharmacol Biochem Behav. 1988;30:107–113.
41. Nakamura K, Kurasawa M, Tanaka Y. Apomorphine-induced hypoattention in rats and reversal of the choice performance impairment by aniracetam. Eur J Pharmacol.

1998;342:127–138.
42. Fernandez-Tome MP, Sanchez-Blazquez P, del Rio J. Impairment by apomorphine of one-trial passive avoidance learning in mice: the opposing roles of the dopamine and noradrenaline systems. Psychopharmacology . 1979;61:43–47.
43. Cabib S, Puglisi-Allegra S, Oliverio A. Chronic stress enhances apomorphine-induced stereotyped behavior in mice: involvement of endogenous opioids. Brain Res. 1984;298:138–140.
44. Goldschmidt PL, Frances H, Simon P. Stimulation of beta-adrenergic receptors and spontaneous motor activity in mice. Pharmacol Biochem Behav. 1984;21:177–180.
45. Weed MR, Gold LH. The effects of dopaminergic agents on reaction time in rhesus monkeys. Psychopharmacology . 1998;137:33–42.
46. Filipowicz ALS, Levine J, Piasini E, Tavoni G, Kable JW, Gold JI. The complexity of model-free and model-based learning strategies. 2019. 2019. https://doi.org/10.1101/2019.12.28.879965.
47. Kool W, McGuire JT, Rosen ZB, Botvinick MM. Decision making and the avoidance of cognitive demand. J Exp Psychol Gen. 2010;139:665–682.
48. Floresco SB. Prefrontal dopamine and behavioral flexibility: shifting from an 'inverted-U' toward a family of functions. Front Neurosci. 2013;7:62.
49. Ebitz RB, Tu JC, Hayden BY. Rules warp feature encoding in decision-making circuits. PLoS Biol. 2020;18:e3000951.
50. Bangasser DA, Wiersielis KR, Khantsis S. Sex differences in the locus coeruleus-norepinephrine system and its regulation by stress. Brain Res. 2016;1641:177–188.
51. Pearson JM, Hayden BY, Raghavachari S, Platt ML. Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. Curr Biol. 2009;19:1532–1537.
52. Ishii S, Yoshida W, Yoshimoto J. Control of exploitation-exploration meta-parameter in reinforcement learning. Neural Netw. 2002;15:665–687.
53. Jepma M, Nieuwenhuis S. Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory. J Cogn Neurosci. 2011;23:1587–1596.
54. Katahira K. The statistical structures of reinforcement learning with asymmetric value updates. J Math Psychol. 2018;87:31–45.
55. Gershman SJ. Empirical priors for reinforcement learning models. J Math Psychol. 2016;71:1–6.
56. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A. 2007;104:16311–16316.
57. Wilson R, Collins A. Ten simple rules for the computational modeling of behavioral data. 2019. 2019.
58. Economides M, Kurth-Nelson Z, Lübbert A, Guitart-Masip M, Dolan RJ. Model-Based Reasoning in Humans Becomes Automatic with Training. PLoS Comput Biol. 2015;11:e1004463.
59. Barto AG. Adaptive critics and the basal ganglia. Models of Information Processing in the Basal Ganglia. 1995;382:215–232.
60. Montague PR, Dayan P, Person C, Sejnowski TJ. Bee foraging in uncertain environments using predictive hebbian learning. Nature. 1995;377:725–728.
61. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems

based on predictive Hebbian learning. J Neurosci. 1996;16:1936–1947.
62. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275:1593–1599.
63. Rutledge RB, Lazzaro SC, Lau B, Myers CE, Gluck MA, Glimcher PW. Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. J Neurosci. 2009;29:15104–15114.
64. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004;306:1940–1943.
65. Beeler JA, Cools R, Luciana M, Ostlund SB, Petzinger G. A kinder, gentler dopamine… highlighting dopamine's role in behavioral flexibility. Front Neurosci. 2014;8:4.
66. Eisenegger C, Naef M, Linssen A, Clark L, Gandamaneni PK, Müller U, et al. Role of dopamine D2 receptors in human reinforcement learning. Neuropsychopharmacology. 2014;39:2366–2375.
67. Lee E, Seo M, Dal Monte O, Averbeck BB. Injection of a dopamine type 2 receptor antagonist into the dorsal striatum disrupts choices driven by previous outcomes, but not perceptual inference. J Neurosci. 2015;35:6298–6306.
68. Beeler JA, Daw N, Frazier CRM, Zhuang X. Tonic dopamine modulates exploitation of reward learning. Front Behav Neurosci. 2010;4:170.
69. Ebitz RB, Sleezer BJ, Jedema HP, Bradberry CW, Hayden BY. Tonic exploration governs both flexibility and lapses. PLoS Comput Biol. 2019;15:e1007475.
70. Sescousse G, Ligneul R, van Holst RJ, Janssen LK, de Boer F, Janssen M, et al. Spontaneous eye blink rate and dopamine synthesis capacity: preliminary evidence for an absence of positive correlation. Eur J Neurosci. 2018;47:1081–1086.
71. Ebitz RB, Moore T. Both a Gauge and a Filter: Cognitive Modulations of Pupil Size. Front Neurol. 2018;9:1190.
72. Findling C, Skvortsova V, Dromnelle R, Palminteri S, Wyart V. Computational noise in reward-guided learning drives behavioral variability in volatile environments. Nat Neurosci. 2019;22:2066–2077.
73. Chakroun K, Mathar D, Wiehler A, Ganzer F, Peters J. Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. Elife. 2020;9.
74. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 2006;442:1042–1045.
75. Orsini CA, Willis ML, Gilbert RJ, Bizon JL, Setlow B. Sex differences in a rat model of risky decision making. Behav Neurosci. 2016;130:50–61.
76. Kane GA, Vazey EM, Wilson RC, Shenhav A, Daw ND, Aston-Jones G, et al. Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. Cogn Affect Behav Neurosci. 2017;17:1073–1083.
77. Joshi S, Gold JI. Pupil Size as a Window on Neural Substrates of Cognition. Trends Cogn Sci. 2020;24:466–480.
78. Iordanova MD, Yau JO-Y, McDannald MA, Corbit LH. Neural substrates of appetitive and aversive prediction error. Neurosci Biobehav Rev. 2021;123:337–351.
79. Mäki-Marttunen V, Andreassen OA, Espeseth T. The role of norepinephrine in the pathophysiology of schizophrenia. Neurosci Biobehav Rev. 2020;118:298–314.
80. Pinos H, Collado P, Rodríguez-Zafra M, Rodríguez C, Segovia S, Guillamón A. The development of sex differences in the locus coeruleus of the rat. Brain Res Bull. 2001;56:73–78.

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