Cognition and Behavior

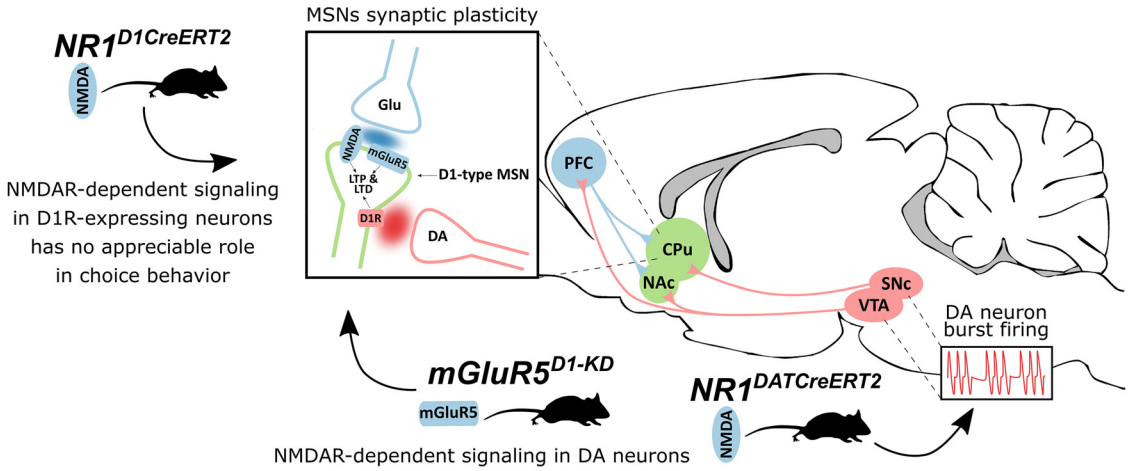
Selective Effects of the Loss of NMDA or mGluR5 Receptors in the Reward System on Adaptive Decision-Making

Przemysław Eligiusz Cieślak,[1] Woo-Young Ahn,[2] Rafał Bogacz,[3] and Jan Rodriguez Parkitna[1]

DOI:http://dx.doi.org/10.1523/ENEURO.0331-18.2018

[1]Department of Molecular Neuropharmacology, Institute of Pharmacology of the Polish Academy of Sciences, 31-343, Krakow, Poland, [2]Department of Psychology, Seoul National University, Seoul 08826, Korea, and [3]MRC Brain Networks Dynamics Unit, Nuffield Department of Clinical Neurosciences, Oxford University, John Radcliffe Hospital, Oxford OX3 9DU, United Kingdom

Visual Abstract



Selecting the most advantageous actions in a changing environment is a central feature of adaptive behavior. The midbrain dopamine (DA) neurons along with the major targets of their projections, including dopaminoceptive neurons in the frontal cortex and basal ganglia, play a key role in this process. Here, we investigate the consequences of a

Significance Statement

We investigated the role of glutamate signaling in the reward system of the brain in adaptive decision-making. We used genetically modified mice with a disruption of glutamate signaling that was caused by the deletion of glutamate receptors in dopamine-producing and dopamine-sensitive neurons. When mutant mice were offered a choice between two alternatives with varying chances of being rewarded, the mutations decreased the probability of selecting the more often rewarded alternative, and the likelihood of repeating a previously rewarded choice. Moreover, mutant animals were much slower in performing choices. Our results show that when glutamate signaling in the reward system is disrupted, it causes an impairment in decision-making by increasing randomness and reducing the speed of the decision-making process.

selective genetic disruption of NMDA receptor and metabotropic glutamate receptor 5 (mGluR5) in the DA system on adaptive choice behavior in mice. We tested the effects of the mutation on performance in the probabilistic reinforcement learning and probability-discounting tasks. In case of the probabilistic choice, both the loss of NMDA receptors in dopaminergic neurons or the loss mGluR5 receptors in D_1 receptor-expressing dopaminoceptive neurons reduced the probability of selecting the more rewarded alternative and lowered the likelihood of returning to the previously rewarded alternative (win-stay). When observed behavior was fitted to reinforcement learning models, we found that these two mutations were associated with a reduced effect of the expected outcome on choice (i.e., more random choices). None of the mutations affected probability discounting, which indicates that all animals had a normal ability to assess probability. However, in both behavioral tasks animals with targeted loss of NMDA receptors in dopaminergic neurons or mGluR5 receptors in D_1 neurons were significantly slower to perform choices. In conclusion, these results show that glutamate receptor-dependent signaling in the DA system is essential for the speed and accuracy of choices, but at the same time probably is not critical for correct estimation of probable outcomes.

Key words: decision-making; dopamine; glutamate receptors; mouse behavior; reinforcement learning

Introduction

Midbrain dopamine (DA) neurons originate from the ventral tegmental area and substantia nigra and, along with the major targets of their projections, including dopaminoceptive neurons in the frontal cortex and basal ganglia, play a central role in the organization of adaptive behavior (Berridge and Robinson, 1998; Wise, 2004; Floresco and Magyar, 2006; Salamone and Correa, 2012). In rodents and nonhuman primates, the burst firing of midbrain DA neurons and the subsequent phasic release of DA encode reward prediction error (Schultz et al., 1997; Bayer and Glimcher, 2005; Hart et al., 2014). This error in reward expectation is a signal of the need to modify synaptic plasticity at corticostriatal synapses and update the action values stored by striatal neurons (Reynolds et al., 2001; Samejima et al., 2005; Lee et al., 2012). In this way, the DA system provides a neural substrate for reinforcement learning mechanisms underlying decision-making and action selection (Glimcher, 2011; Schultz, 2015). It should be noted though that the role of DA extends beyond reinforcement learning, as it is also involved in the regulation of motivation and vigor as well as the performance of instrumental behavior (Salamone and Correa, 2012; Shiner et al., 2012).

The activity and plasticity in the DA system are largely dependent on excitatory glutamatergic transmission. Glutamatergic inputs activate NMDA receptors and drive the burst firing in DA neurons (Overton and Clark, 1992; Chergui et al., 1993), phasic DA release (Sombers et al., 2009; Wickham et al., 2013), and induction of long-term potentiation onto the dopaminergic neurons underlying cue–reward learning (Stuber et al., 2008; Harnett et al., 2009). Moreover, NMDA receptors and metabotropic glutamate receptor 5 (mGluR5) are crucial for the induction of synaptic and structural plasticity in dopaminoceptive striatal medium spiny neurons (Calabresi et al., 2007; Shen et al., 2008; Surmeier et al., 2009; Yagishita et al., 2014). Altogether, these observations indicate that glutamate-dependent signaling is crucial for DA-mediated reinforcement. However, in most studies, the observations are based on correlations and *in vitro* measurements; therefore, the causality or degree of contribution remains uncertain.

A more direct approach for testing the role of glutamate-dependent signaling in reinforcement learning is the use of genetically modified mice with an inactivation of glutamate receptors in DA or dopaminoceptive neurons. Such models have been generated and generally observed to result in impairments in tasks involving instrumental and pavlovian learning, confirming that a disruption in glutamate-dependent signaling in the DA system is sufficient to cause an impairment in reward-based learning (Zweifel et al., 2009; Novak et al., 2010; Parker et al., 2010, 2011; Beutler et al., 2011; Wang et al., 2011; James et al., 2015). However, most experiments were conducted using paradigms in which only a single lever or conditioned stimulus was reinforced. Therefore, a crucial aspect of adaptive decision-making (i.e., choosing among competing courses of action in a changing environment) was not comprehensively addressed in those studies.

Here, we sought to determine the contribution of glutamate receptor-dependent signaling in DA and dopaminoceptive neurons to adaptive decision-making. We used mice with cell type-specific, tamoxifen-inducible inactivation of NMDA receptors in DA and D_1 receptor-expressing neurons (Engblom et al., 2008; Jastrzębska et al., 2016; Sikora et al., 2016) and animals with a knockdown of mGluR5 receptors in D_1 neurons (Novak et al., 2010; Rodriguez Parkitna et al., 2013). The animals were tested using a probabilistic reinforcement learning task, in which the mouse is required to estimate the expected value of two alternatives associated with different reward proba-

Received September 22, 2017; accepted June 3, 2018; First published July 13, 2018.

The authors declare no competing financial interests.

Author contributions: P.E.C. and J.R.P. designed research; P.E.C. performed research; W.-Y.A. contributed unpublished reagents/analytic tools; P.E.C., W.-Y.A., R.B., and J.R.P. analyzed data; P.E.C., W.-Y.A., R.B., and J.R.P. wrote the paper.

This work was supported by the Polish National Science Centre Grant PRELUDIUM (2014/15/N/NZ4/00761). P.E.C. is a recipient of the ETIUDA scholarship from the Polish National Science Centre (2016/20/T/NZ4/00503). R.B. was supported by the Medical Research Council (Grant MC_UU_12024/5).

Acknowledgments: We thank Dr. Nii Addy for helpful comments.

Correspondence should be addressed to Jan Rodriguez Parkitna, Department of Molecular Neuropharmacology, Institute of Pharmacology of the Polish Academy of Sciences, Smętna 12, 31-343 Krakow, Poland. E-mail: Jan.Rodriguez@if-pan.krakow.pl[MAIL]

DOI:http://dx.doi.org/10.1523/ENEURO.0331-18.2018

bilities by trial and error. This task was followed by a probability-discounting task in which the animal is required to choose between two options that provide rewards that differ in magnitude (small vs large) and probability (certain vs uncertain).

Materials and Methods

Animals

The following three strains of genetically modified mice were used in the study: $NR1^{DATCreERT2}$ mice, which had an inducible deletion of the NR1 subunit of the NMDA receptor in DA transporter (DAT)-expressing neurons (Engblom et al., 2008; Jastrzębska et al., 2016); $NR1^{D1CreERT2}$ animals, which had an inducible loss of the NR1 subunit of the NMDA receptor in D_1 receptor-expressing neurons (Sikora et al., 2016); and $mGluR5^{KD\text{-}D1}$ mice, which had a selective knockdown of the mGluR5 receptor in D_1-expressing neurons (Novak et al., 2010; Rodriguez Parkitna et al., 2013). All strains were bred to be congenic with the C57BL/6N strain. Genotyping was performed as previously described. The animals were housed two to five animals per cage in a room with a controlled temperature at 22 ± 2°C under a 12 h light/dark cycle. Unless otherwise indicated, the mice had *ad libitum* access to tap water and standard rodent laboratory chow.

Regarding the CreERT2-dependent mutations, the recombination was induced in adult animals at the age of 8–10 weeks using tamoxifen treatment. Tamoxifen (Sigma-Aldrich) was dissolved in sunflower oil, filtered through a 0.22 μm membrane, and injected intraperitoneally once a day for 5 consecutive days at a dose of 100 mg/kg and a volume of 5 μl/g. The genotype of the mutant mice was [Tg/0; flox/flox], and the genotype of the control animals was [0/0; flox/flox]. All tamoxifen-treated animals were allowed to rest for at least 3 weeks before the start of the behavioral procedures. Regarding $mGluR5^{KD\text{-}D1}$, no induction was necessary, and the expression of the transgene was initiated when the D_1 promoter became active during late development. The genotype of the mutant $mGluR5^{KD\text{-}D1}$ animals was [Tg/0], and the genotype of their respective controls was [0/0].

Only male mice were used in the study. The mean ages and weights of the cohorts of animals used in the experiments were as follows: 16.25 ± 1.05 weeks and 25.6 ± 0.85 × *g* for the $NR1^{DATCreERT2}$ mice and 16.57 ± 1.15 weeks and 29.43 ± 0.62 × *g* for their respective controls; 18.33 ± 0.94 weeks and 26.39 ± 1.21 × *g* for the $NR1^{D1CreERT2}$ mice and 19.33 ± 1.08 weeks and 27.85 ± 1.37 × *g* for their controls; and 13.38 ± 1.31 weeks and 25.8 ± 1.1 × *g* for the $mGluR5^{KD\text{-}D1}$ mice and 13.56 ± 1.12 weeks and 24.98 ± 1.02 × *g* for their controls. The same cohorts of animals were used in the probabilistic reinforcement learning and probability-discounting tasks.

Behavioral procedures

Water deprivation

A week before the behavioral testing, water consumption was limited to 1–1.5 ml/d, and this water restriction schedule was maintained for the duration of the experiments. The mice were trained 5–7 d/week, and their body weight was monitored daily. The water restriction was lessened if the mice fell to <85% of their body weight from the beginning of the deprivation.

Apparatus

The experiments were performed using mouse operant chambers (ENV-307W-CT, Med Associates) enclosed in cubicles that were equipped with a fan to provide ventilation and mask extraneous noise. Each chamber was equipped with a dual cup liquid receptacle, a nose-poke port containing a cue light located on each side of a liquid receptacle, and a house light located on the wall opposite to the liquid receptacle. Saccharin-flavored water (0.01% w/v saccharin; Sigma-Aldrich) was delivered into an individual cup by an infusion pump (PHM-100, Med Associates) connected to the liquid receptacle via a silicone tube. The amount of fluid delivered (reward size) was dependent on the duration of the infusion.

Training

First, the mice were placed in the operant chamber for 30 min, during which 20 μl of water were delivered into the receptacle in 60 s intervals. This procedure allowed the animals to become familiar with the chamber and liquid reward. On subsequent days, the mice were trained under a continuous reinforcement schedule and were rewarded with 10 μl of water after poking their noses into the active port (with the cue-light on). The other port was inactive. The nose pokes in the inactive port were recorded but had no consequences. The port assignment was counterbalanced, and the animals were trained until they reached the criterion of 60 rewarded responses in 40 min, which occurred first in one port and then in the other port in a subsequent session. This training was followed by additional training during which the left and right ports were active once in every pair of trials, and the order within the pair was random. These sessions ended when an animal completed 100 trials or 60 min elapsed, whichever came first. There was no limit to the trial duration, and each trial ended when a nose poke in the active port resulted in the delivery of a reward, followed by a 5 s intertrial interval (ITI). The animals had to complete at least 85 trials. Finally, the mice underwent omission training, which was similar to the training described above with two exceptions. First, the trial number was increased to 160. Second, responding in an active poke resulted in a 50% chance of reward omission. Reward omission was signaled by switching on the house light for the duration of the ITI. The animals had to complete at least 120 trials.

Probabilistic reinforcement learning task

In this task, the nose-poke ports were randomly assigned reward probabilities of 80% or 20% (Fig. 1*A*). During each session, the reward probabilities were reversed after 60 trials. Thus, to maximize the long-term sum of the rewards, the mouse had to select the alternative with the higher success probability and adapt its choices to the changes in the reward contingencies. There was no limit to the trial duration, and the session ended when the animal completed 120 trials or 60 min elapsed. Rewarded choices resulted in the delivery of 10 μl of water, followed by a 5 s ITI. Unrewarded choices

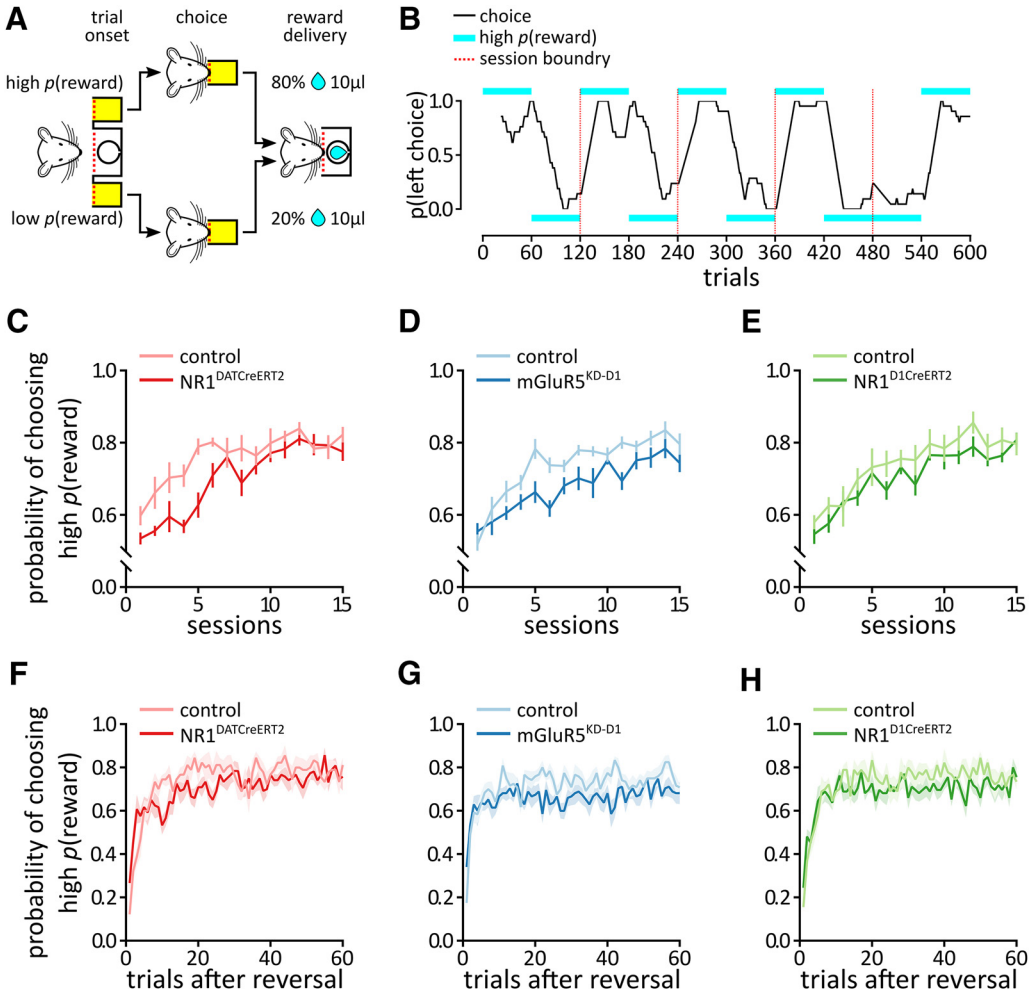


Figure 1. The probabilistic reinforcement learning task. *A*, Schematic representation of the probabilistic reinforcement learning task. The animal could make a nose-poke in one of two ports. Following a nose-poke, water could have been delivered with the probability depending on the chosen port. The nose-poke ports were randomly assigned 80% or 20% reward probabilities. During each session, the reward probabilities were reversed after 60 trials. *B*, An example the choice behavior of a mouse in 600 trials (sessions 6–10). The black line shows the probability of choosing the left side (data smoothed with the 21 point moving average). The cyan bars indicate the side with the higher probability of reward delivery. The red dashed line indicates session boundaries. *C–H*, Probability of selecting the alternative with the higher reward probability by the $NR1^{DATCreERT2}$ (mutant, $n = 6$; control, $n = 8$; *C*, *F*), $mGluR5^{KD\text{-}D1}$ (mutant, $n = 8$; control, $n = 9$; *D*, *G*), and $NR1^{D1CreERT2}$ (mutant, $n = 6$; control, $n = 9$; *E*, *H*) strains. *C–E*, Session-by-session analysis; data were collapsed across trials. *F–H*, Trial-by-trial analysis; data were collapsed across sessions. Data are represented as the mean ± SEM.

were signaled by turning on the house light for the duration of the ITI. The animals were trained in this task for 15 sessions.

Probability-discounting task

In this task, one nose-poke port was associated with the delivery of a small reward (10 µl), while the other nose-poke port was associated with the delivery of a large reward (20 µl). Each session consisted of 20 forced trials, followed by 40 free choice trials (see Fig. 5*A*). During the forced trials, only one port was active, whereas during the free choice trials, both ports were active. Once the preference for the large reward was stabilized, the probability of its delivery gradually decreased to 75%, 50%, or 25% during subsequent blocks of four to five sessions. Simultaneously, the small reward was always available at a 100% probability. The trials were separated by a 5 s ITI, and unrewarded choices were signaled by turning on the house light for the duration of the ITI.

Statistical analysis

A script written in R was used to parse the data files that were generated during the behavioral experiments. All statistical analyses were conducted using GraphPad Prism 7 (GraphPad Software) and R software. Statistical significance was estimated using an ANOVA, followed by a Bonferroni *post hoc* test or a Student's *t* test, as appropriate. The results were considered significant at $\alpha = 0.05$. One animal from the control group in the $NR1^{DATCreERT2}$ strain was classified as an outlier (Grubb's test) in choice reaction time measures (in both tasks) and was excluded from all analyses. Two animals (one from the $NR1^{DATCreERT2}$ strain and one control mouse from the

NR1^{D1CreERT2} strain) showed no preference for the freely available large reward in the probability-discounting task (0.5 ± 0.5% and 1.5 ± 0.6%, respectively) and were excluded from the analysis of this task, to avoid the misinterpretation of the effect of discounting. Confidence intervals (CIs) for *post hoc* comparisons are listed in Table 1.

Computational modeling

We fitted three reinforcement learning models to trial-by-trial choice data of the probabilistic reinforcement learning task, which are all based on the Rescorla-Wagner model (Rescorla and Wagner, 1972), but include additional features. Model 1 assumes that animals learn with different rates when the reward prediction error is positive and negative (den Ouden et al., 2013). Model 2 assumes that the animals have learned that entering only one of the ports results in a high reward probability, so in this model after choosing one option, the expected rewards for both options are modified in opposite directions (Gläscher et al., 2009). Model 3 integrates models 1 and 2, so it includes separate learning rates for positive and negative prediction errors of the chosen option and updates the unchosen option using the fictitious learning component of model 2.

As model 3 is the most general, we start with its description, and then present how it can be simplified to give models 1 and 2. In model 3, the expected value of the chosen ($V_{c,t}$) and unchosen ($V_{uc,t}$) options are updated as follows on each trial t. If prediction error on trial t ($PE_t = r_t - V_{c,t}$) is ≥0, expected values of chosen and unchosen options are updated with learning rate $\eta_+ (0 \le \eta_+ \le 1)$, as follows:

$$V_{c,t+1} = V_{c,t} + \eta_+ \cdot (r_t - V_{c,t}) \quad (1)$$

$$V_{uc,t+1} = V_{uc,t} + \eta_+ \cdot (-r_t - V_{uc,t}). \quad (2)$$

Note that the unchosen option is updated with a fictitious prediction error ($PE_t = -r_t - V_{uc,t}$) following the study by Gläscher et al. (2009). If PE_t is <0, the expected values of chosen and unchosen options are updated with learning rate $\eta_- (0 \le \eta_- \le 1)$:

$$V_{c,t+1} = V_{c,t} + \eta_- \cdot (r_t - V_{c,t}) \quad (3)$$

$$V_{uc,t+1} = V_{uc,t} + \eta_- \cdot (-r_t - V_{uc,t}). \quad (4)$$

In the simulations, r_t is set to 1 if reward is received on trial t, or to −1 if it is omitted. Choice probabilities are computed based on the expected values as follows. If A and B refer to the two options of the probabilistic reinforcement learning task and $p_{t+1}(A)$ refers to the probability of choosing the option A on trial $t + 1$, then:

$$p_{t+1}(A) = \frac{1}{1 + e^{-\beta \cdot (V_{A,t} - V_{B,t})}}. \quad (5)$$

Here, $\beta (0 \le \beta)$ is the inverse temperature parameter, which governs the degree of exploitation and exploration (i.e., low and high values of β indicate more exploration and exploitation, respectively). In summary, model 3 has three free parameters: η_+ (learning rate for positive PE), η_- (learning rate for negative PE) and β (inverse temperature). If we set $\eta_+ = \eta_- = \eta$, the model becomes model 2, which has two free parameters: η (learning rate) and β (inverse temperature). If we only update the values of chosen options using Equations 1 and 3 (but not use Equations 2 and 4), the model becomes model 1, which also has three free parameters: η_+ (learning rate for positive PE), η_- (learning rate for negative PE), and β (inverse temperature).

We fitted the three models using hierarchical Bayesian analysis (HBA), which pools information across individuals and allows us to capture both individual differences and commonalities across subjects in a reliable way (Shiffrin et al., 2008; Ahn et al., 2011; Lee, 2011). To perform HBA, we used the hBayesDM package (Ahn et al., 2017), which is an R package that offers HBA of various computational models and tasks using the Stan software (Carpenter et al., 2017). The hBayesDM functions of models 1–3 are *prl_rp*, *prl_fictitious_woa*, and *prl_fictitious_rp_woa*, respectively. All source codes and Bayesian model formulation are available in its GitHub repository: https://github.com/CCS-Lab/hBayesDM. We performed model comparisons and identified a best-fitting model using leave-one-out cross-validation information criterion (LOOIC). To compute LOOIC for a given model we used the *loo* R package, which computes leave-one-out predictive density using Pareto smoothed importance sampling (Vehtari et al., 2017). The LOOIC inherently penalizes model complexity, as an overly complicated model will perform poorly on unseen data than a simpler model. It also has an advantage over other measures designed to prevent overfitting by overly complex model (like Akaike or Bayesian information criterion) in that it measures the overfitting directly.

Simulation Analysis

To test whether the best-fitting model can describe the observed data well, we performed simulation analysis as previously described (Ahn et al., 2008; Steingroever et al., 2014). Briefly, by using estimated individual parameters alone (without access to trial-by-trial choice history), we generated simulated agents and computed their win-stay and lose-shift (switching to the alternative choice when the preceding response yielded no reward) probabilities. When we generated simulated data, for each group and condition, we used its total number of trials and subjects of the real data. Then, we simulated choices on the probabilistic reinforcement learning task using estimated individual parameters (individual posterior means) of each simulated agent for the whole trajectory (i.e., 1800 trials) using customized R codes.

Results

Performance in the probabilistic reinforcement learning task

The animals were tested in a probabilistic reinforcement learning task in which they could choose between two alternatives with either an 80% or 20% chance of being

Table 1. Statistical table

Figure	Data structure	Type of test	95% CIs or 95% HDIs
Figure 2A $_{\eta+}$	Assumed normal distribution	Hyperposterior distribution	(−0.3601, 0.0779)
Figure 2A $_{\eta-}$	Assumed normal distribution	Hyperposterior distribution	(−0.2952, 0.129)
Figure 2A $_{\beta}$	Assumed normal distribution	Hyperposterior distribution	(−0.461, −0.1018)
Figure 2B $_{\eta+}$	Assumed normal distribution	Hyperposterior distribution	(−0.3081, 0.2367)
Figure 2B $_{\eta-}$	Assumed normal distribution	Hyperposterior distribution	(−0.2532, 0.4388)
Figure 2B $_{\beta}$	Assumed normal distribution	Hyperposterior distribution	(−0.5163, −0.1429)
Figure 2C $_{\eta+}$	Assumed normal distribution	Hyperposterior distribution	(−0.3631, 0.0847)
Figure 2C $_{\eta-}$	Assumed normal distribution	Hyperposterior distribution	(−0.2793, 0.2792)
Figure 2C $_{\beta}$	Assumed normal distribution	Hyperposterior distribution	(−0.4919, 0.2115)
Figure 3A $_{\text{win-stay}}$	Assumed normal distribution	Two-tailed *t* test	(−0.126, −0.03798)
Figure 3A $_{\text{lose-shift}}$	Assumed normal distribution	Two-tailed *t* test	(−0.05501, 0.1081)
Figure 3B $_{\text{win-stay}}$	Assumed normal distribution	Two-tailed *t* test	(−0.1675, −0.03254)
Figure 3B $_{\text{lose-shift}}$	Assumed normal distribution	Two-tailed *t* test	(−0.01564, 0.1316)
Figure 3C $_{\text{win-stay}}$	Assumed normal distribution	Two-tailed *t* test	(−0.1263, 0.01433)
Figure 3C $_{\text{lose-shift}}$	Assumed normal distribution	Two-tailed *t* test	(−0.01022, 0.1521)
Figure 3D $_{\text{win-stay}}$	Assumed normal distribution	Two-tailed *t* test	(−0.126, −0.03624)
Figure 3D $_{\text{lose-shift}}$	Assumed normal distribution	Two-tailed *t* test	(−0.0746, 0.04841)
Figure 3E $_{\text{win-stay}}$	Assumed normal distribution	Two-tailed *t* test	(−0.1797, −0.03847)
Figure 3E $_{\text{lose-shift}}$	Assumed normal distribution	Two-tailed *t* test	(−0.02373, 0.05712)
Figure 3F $_{\text{win-stay}}$	Assumed normal distribution	Two-tailed *t* test	(−0.1268, 0.01674)
Figure 3F $_{\text{lose-shift}}$	Assumed normal distribution	Two-tailed *t* test	(−0.02587, 0.1276)
Figure 4B $_{\text{lose}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−0.151, −9.144)
Figure 4B $_{\text{win}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−5.636, −14.629)
Figure 4B $_{\text{control}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(0.535, −9.078)
Figure 4B $_{\text{mutant}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−5.594, −13.919)
Figure 4C	Assumed normal distribution	Two-tailed *t* test	(0.07629, 0.2383)
Figure 4E $_{\text{lose}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(1.211, −5.670)
Figure 4E $_{\text{win}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−2.656, −9.537)
Figure 4E $_{\text{control}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−4.688, −11.769)
Figure 4E $_{\text{mutant}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−8.758, −15.433)
Figure 4F	Assumed normal distribution	Two-tailed *t* test	(0.03127, 0.2603)
Figure 4H $_{\text{lose}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(3.862, −5.833)
Figure 4H $_{\text{win}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(2.707, −6.988)
Figure 4H $_{\text{control}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(1.383, −9.238)
Figure 4H $_{\text{mutant}}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−0.746, −9.418)
Figure 4I	Assumed normal distribution	Two-tailed *t* test	(−0.09834, 0.06706)
Figure 5B $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(44.207, −33.477)
Figure 5B $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(42.318, −35.366)
Figure 5B $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(46.199, −31.485)
Figure 5B $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(23.622, −54.062)
Figure 5C $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(18.145, −26.659)
Figure 5C $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(13.173, −31.631)
Figure 5C $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(18.416, −26.388)
Figure 5C $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(18.624, −26.179)
Figure 5D $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(25.957, −20.935)
Figure 5D $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(18.868, −28.024)
Figure 5D $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(1.613, −45.279)
Figure 5D $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(19.101, −27.791)
Figure 6A forced $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−8.519, −16.751)
Figure 6A forced $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−7.398, −15.630)
Figure 6A forced $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−6.524, −14.756)
Figure 6A forced $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−4.346, −12.578)
Figure 6A free $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−6.166, −15.530)
Figure 6A free $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−2.561, −11.925)
Figure 6A free $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−1.947, −11.312)
Figure 6A free $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−1.615, −10.979)
Figure 6B forced $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−2.772, −8.595)
Figure 6B forced $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−1.419, −7.243)
Figure 6B forced $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−0.924, −6.748)
Figure 6B forced $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−1.544, −7.368)
Figure 6B free $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(−0.135, −7.745)
Figure 6B free $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(0.421, −7.189)
Figure 6B free $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(0.512, −7.098)

Figure	Data structure	Type of test	95% CIs or 95% HDIs
Assumed normal distribution	Bonferroni-corrected *t* test	(0.636, −6.974)	
Figure 6C forced $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(2.767, −5.107)
Figure 6C forced $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(3.120, −4.754)
Figure 6C forced $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(3.388, −4.486)
Figure 6C forced $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(2.297, −5.578)
Figure 6C free $_{100\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(2.447, −5.754)
Figure 6C free $_{75\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(3.649, −4.552)
Figure 6C free $_{50\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(4.181, −4.020)
Figure 6C free $_{25\%}$	Assumed normal distribution	Bonferroni-corrected *t* test	(0.952, −7.249)

rewarded with 10 μl of water (Fig. 1*A*). The test consisted of 15 sessions, and each session consisted of 120 trials. The trials were not time limited. The initial assignment of the reward probabilities was random and reversed in the middle of each session. An example of the choice behavior of a mouse over 600 trials (sessions 6–10) is shown in Fig. 1*B*.

All groups, regardless of their genotype, showed a significant increase in the frequency of selecting the more often rewarded alternative over the course of the experiment (Fig. 1*C*: session, $F_{(14,168)} = 17.15$; Fig. 1*D*: session $F_{14,210} = 20.69$; Fig. 1*E*: session, $F_{(14,182)} = 19.17$; all $p < 0.0001$). The NR1^{DATCreERT2} mice chose the alternative with the higher reward probability on a smaller fraction of trials (Fig. 1*C*: genotype, $F_{(1,12)} = 11.50$, $p = 0.0054$). However, this difference was due to initial slower increase in choosing the correct option, and the mutants eventually reached the same performance levels as the control animals (genotype × session: $F_{(14,168)} = 1.90$, $p = 0.0298$). In contrast, in the mGluR5$^{KD\text{-}D1}$ mice, the probability of choosing the alternative with the higher reward probability was consistently lower (Fig. 1*D*: genotype, $F_{(1,15)} = 12.62$, $p = 0.0029$; genotype × session, $F_{(14,210)} = 1.49$, $p = 0.1180$). The choice behavior of the NR1^{D1CreERT2} mice did not differ from that of the controls (Fig. 1*E*: genotype, $F_{(1,13)} = 1.79$, $p = 0.2034$; genotype × session, $F_{(14,182)} = 0.53$, $p = 0.9103$).

Figure 1*F–H* shows the probability of choosing the correct option in the 60 trials after reversal (average based on all sessions). The probability was initially <50%, as mice choose the option that was rewarded more frequently before reversal, but then quickly increased (Fig. 1*F*: trial, $F_{(59,708)} = 12.6$; Fig. 1*G*: trial, $F_{(59,885)} = 7.03$; Fig. 1*H*: trial, $F_{(59,767)} = 9.67$; all $p < 0.0001$). The effects of mutations in Figure 1*F–H* parallel those observed in Figure 1*C–E*. The NR1^{DATCreERT2} mice were initially slower in choosing the alternative with the higher reward probability, but eventually reached the same performance levels as the control animals (Fig. 1*F*: genotype, $F_{(1,12)} = 1.83$, $p = 0.20$; genotype × trial, $F_{(59,708)} = 1.86$, $p = 0.0002$). The mGluR5$^{KD\text{-}D1}$ mice chose the alternative with the higher reward probability less frequently (Fig. 1*G*: genotype, $F_{(1,15)} = 7.55$, $p = 0.015$), and this difference depended on the trial number (genotype × trial, $F_{(59,885)} = 1.43$, $p = 0.02$), but to a lower extent than for the NR1^{DATCreERT2} mice. The choice behavior of the NR1^{D1CreERT2} mice did not differ from that of controls (Fig. 1*H*: genotype $F_{(1,13)} = 3.32$, $p = 0.092$; genotype × trial, $F_{(59,767)} = 1.07$, $p = 0.34$).

Computational modeling results

We tested fits of three reinforcement learning models based on reward prediction error. Table 2 shows the LOOIC scores for the three models compared. For all groups tested, model 3 outperformed the others and had the lowest LOOIC scores by a large margin. Model 3 assumes that animals learn with different rates when the prediction error is positive or negative, and also that mice take the higher-order structure of the task into account, namely that they learn that at a given time only one of the ports gives high reward probability. Thus, in model 3 when unexpected reward is obtained following nose-poke to the left port, the expected reward associated with this port is increased, while the expected reward for the right port is decreased.

A summary of parameters calculated for the best-fitting model is shown in Figure 2*A–C*. For each parameter, we quantified an effect of the mutation by calculating the difference of hyperposterior distributions between mutant and control mice (Ahn et al., 2014), which is summarized as the 95% highest density interval (HDI). The 95% HDI refers to the range of parameter values that span the 95% of the distribution (Kruschke, 2014). If the 95% HDI of the difference is far >0 or <0, it indicates that there is a strong evidence of a group difference. While binary interpretations of 95% HDI should be avoided, it is possible to check whether the 95% HDI excludes 0 for a heuristic judgment of "credible" group differences. As in the case of previous analyses, credible effects of mutations were observed in the NR1^{DATCreERT2} mice (95% HDI = [−0.461, −0.102]) and mGluR5$^{KD\text{-}D1}$ mice (95% HDI = [−0.516, −0.143]). We found that the mutation in the NR1^{DATCreERT2} and mGluR5$^{KD\text{-}D1}$ strains affected the inverse temperature (β) parameter and mutant mice make more random rather than value-driven choices. However, the mutation did not

Table 2. Model comparisons using the LOOIC

Group	Model 1	Model 2	**Model 3**
NR1^{DATCreERT2}	16,167.7	14,203.5	**14,129.5**
Control	9545.9	8304.4	**8256.2**
mGluR5$^{KD\text{-}D1}$	18,640.8	17,888.8	**17,643.2**
Control	14,216.2	12,743.5	**12,557.1**
NR1^{D1CreERT2}	17,247.8	16,011.6	**15,801.0**
Control	10,449.9	8753.4	**8687.3**

Lower values of LOOIC indicate better model fits. The best performing model is highlighted with bold type; model 3 outperformed other models in all groups. Model 1, Separate learning rates for positive and negative reward PE; model 2, a single learning rate for PE and fictitious updating for the unchosen option; model 3, separate learning rates for positive and negative PE and fictitious updating for the unchosen option.

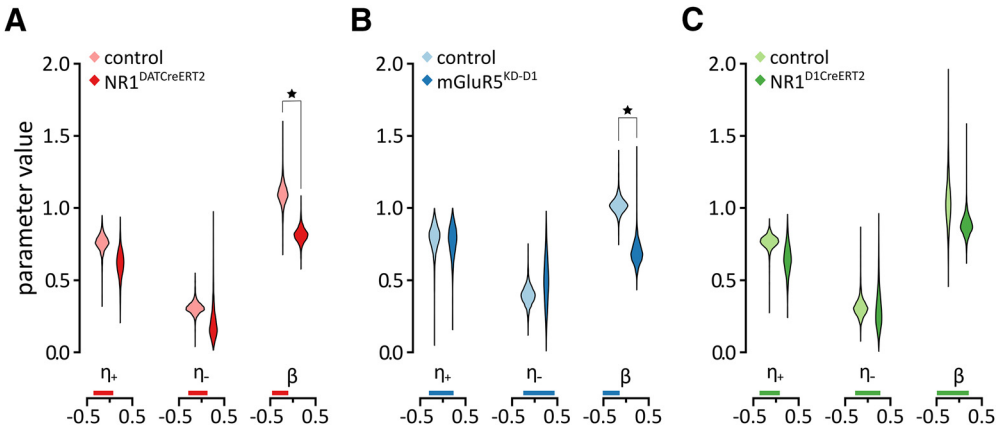


Figure 2. Computational modeling results. *A–C*, Density plots of posterior group parameter distributions with the best model (model 3) for the NR1^{DATCreERT2} (*A*), mGluR5^{KD-D1} (*B*), and NR1^{D1CreERT2} (*C*) strains. Credible differences are marked with stars, and vertical bars below the plots show 95% HDI ranges.

cause a credible difference in the case of the NR1^{D1CreERT2} mice (95% HDI = [−0.492, 0.212]).

In agreement with the analysis of learning behavior of the NR1^{DATCreERT2} group (Fig. 1*F*), the means of posterior distribution of the learning rates for this group were lower than those of controls (Fig. 2*A*). However, unexpectedly, this effect was not credible (95% HDI = [−0.360, 0.078] for the reward learning rate; 95% HDI = [−0.295, 0.129] for the punishment learning rate). We did not observe any other credible effects of any of the mutations on learning rates. Another interesting observation was that in all groups, learning rates tend to be higher for positive than negative outcomes. Such a relationship between the learning rates has been observed before in a probabilistic choice task, and was proposed to arise because the animals might have learned that one option gives a higher reward on average, so a single reward omission may just be noise and should not change the behavior (Grogan et al., 2017). In summary, the computational modeling indicated that mutations significantly affected only the parameter influencing the preference for the alternative with a higher expected outcome. Additionally, the behavior in general was most consistent with models that included updates of the expected value of the nonselected alternative.

Effects of prior outcomes on choice

To further assess the influence of previous outcomes on subsequent choices, we calculated the probabilities of repeating the same choice when the previous response was rewarded (win-stay) and switching to the alternative choice when the preceding response yielded no reward (lose-shift; Fig. 3*A–C*). The NR1^{DATCreERT2} and mGluR5^{KD-D1} mice were significantly less likely to repeat the previously rewarded choice than the control animals, whereas neither mutation affected the lose-shift ratio (Fig. 3*A*: win-stay, $t_{(12)} = 4.059$, $p = 0.0016$; lose-shift: $t_{(12)} = 0.7093$, $p = 0.4917$; Fig. 3*B*: win-stay, $t_{(15)} = 3.159$, $p = 0.0065$; lose-shift, $t_{(15)} = 1.679$, $p = 0.1139$). No significant effect of genotype on win-stay or lose-shift responding was observed in the NR1^{D1CreERT2} animals (Fig. 3*C*: win-stay, $t_{(13)} = 1.72$, $p = 0.1091$; lose-shift, $t_{(13)} = 1.888$, $p = 0.0815$).

The overall higher proportion of win-stay than lose-shift trials is in a qualitative agreement with the higher learning rate from positive than from negative feedback (Fig. 2). To test whether the model can quantitatively reproduce the proportions of win-stay trials and lose-shift trials, Figure 3*D–F* shows the simulation performance of model 3 with parameters set to the means of posterior distributions in Figure 2*A–C*. Comparisons of actual (Fig. 3*A–C*) and simulated (Fig. 3*D–F*) behavioral performance revealed that our model indeed describes observed data very well. Consistent with actual data, simulated NR1^{DATCreERT2} and mGluR5^{KD-D1} mice were significantly less likely to repeat the previously rewarded choice than the control animals (win-stay), but this was not observed in NR1^{D1CreERT2} simulated mice (Fig. 3*D*: win-stay, $t_{(12)} = 3.939$, $p = 0.0020$; Fig. 3*E*: win-stay, $t_{(15)} = 3.292$, $p = 0.0049$; Fig. 3*F*: win-stay, $t_{(13)} = 1.657$, $p = 0.1215$). We observed no effect of mutation on lose-shift behavior in any group, which is consistent with actual data (Fig. 3*D*: lose-shift, $t_{(12)} = 0.4638$, $p = 0.6511$; Fig. 3*E*: lose-shift, $t_{(15)} = 0.8803$, $p = 0.3926$; Fig. 3*F*: lose-shift, $t_{(13)} = 1.432$, $p = 0.1757$).

Choice latency

The analysis of the reaction times in the probabilistic reinforcement learning task revealed that the NR1^{DATCreERT2} and mGluR5^{KD-D1} mice required significantly more time to make a choice after the trial onset (Fig. 4*A*: genotype × trial, $F_{(119,1428)} = 0.90$, $p = 0.7764$; genotype, $F_{(1,12)} = 34.89$, $p < 0.0001$; trial, $F_{(119,1428)} = 1.07$, $p = 0.2910$; Fig. 4*D*: genotype × trial, $F_{(119,1785)} = 0.84$, $p = 0.8871$; genotype, $F_{(1,15)} = 10.51$, $p = 0.0055$; trial, $F_{(119,1785)} = 3.62$, $p < 0.0001$). Furthermore, the choice latency was strongly affected by the previous outcome, and the NR1^{DATCreERT2} and mGluR5^{KD-D1} mice spent more time choosing when the previous trial was rewarded (Fig. 4*B*: genotype × outcome, $F_{(1,24)} = 6.15$, $p = 0.0205$; genotype, $F_{(1,24)} = 44.66$; outcome, $F_{(1,24)} = 40.23$; both $p < 0.0001$; Fig. 4*E*: genotype × outcome,

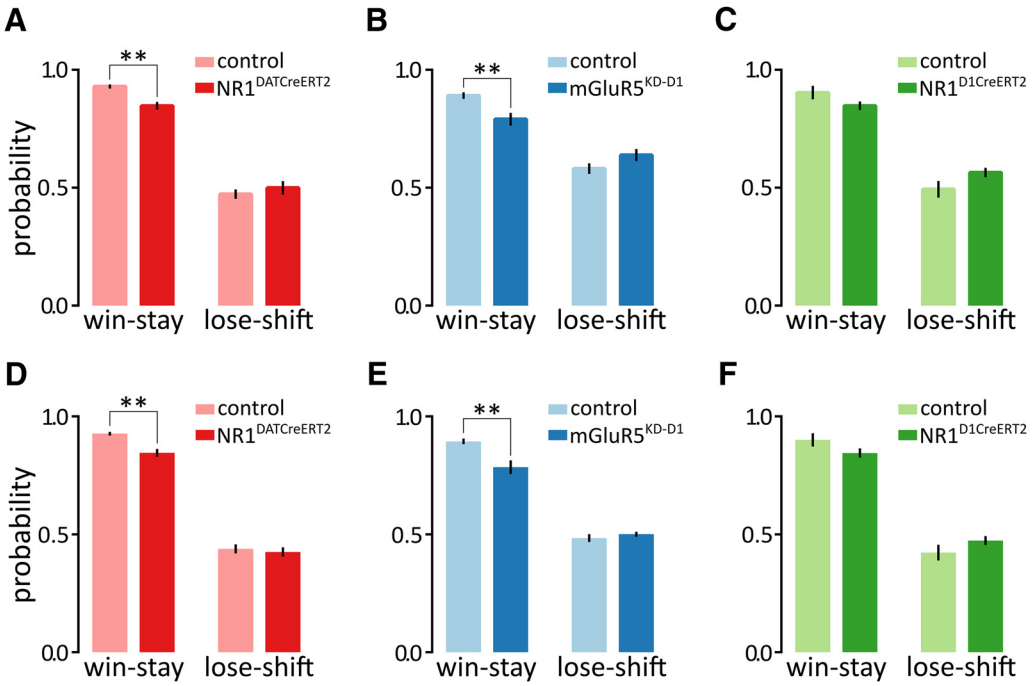


Figure 3. Effects of previous outcomes on choice. *A–C*, Probabilities of repeating the same choice when the previous response was rewarded (win-stay) or switching to an alternative choice when the preceding response yielded no reward (lose-shift) in the $NR1^{DATCreERT2}$ (mutant, $n = 6$; control, $n = 8$; *A*), $mGluR5^{KD\text{-}D1}$ (mutant, $n = 8$; control, $n = 9$; *B*), and $NR1^{D1CreERT2}$ (mutant, $n = 6$; control, $n = 9$; *C*) strains. The probability of win-stay was calculated as the number of times the animal chose the same side as the side chosen during the previously rewarded trial divided by the total number of rewarded trials, while the lose-shift probability was calculated as the number of times the animal changed its choice when the preceding response yielded no reward divided by the total number of unrewarded trials. *D–F*, Simulation performance of the best model (model 3) with respect to mimicking win-stay/lose-shift choice behavior. Data are represented as the mean ± SEM. $^{**}p < 0.01$ (*t* test).

$F_{(1,30)} = 5.04$, $p = 0.0323$; genotype, $F_{(1,30)} = 23.37$; outcome, $F_{(1,30)} = 139.21$; both $p < 0.0001$). In addition, the $NR1^{DATCreERT2}$ and $mGluR5^{KD\text{-}D1}$ mice were slightly slower to collect their reward (Fig. 4*C*: $t_{(12)} = 4.3$, $p = 0.0010$; Fig. 3*F*: $t_{(15)} = 3.242$, $p = 0.0055$). Again, no effect of mutation on decision time or reward latency was observed in the $NR1^{D1CreERT2}$ strain (Fig. 4*G*: genotype × trial, $F_{(119,1547)} = 0.93$, $p = 0.6885$; genotype, $F_{(1,13)} = 1.45$, $p = 2499$; trial, $F_{(119,1547)} = 3.16$, $p < 0.0001$; Fig. 4*H*: genotype × outcome, $F_{(1,26)} = 0.23$, $p = 0.6347$; genotype, $F_{(1,26)} = 1.70$, $p = 0.2043$; outcome, $F_{(1,26)} = 14.08$, $p = 0.0009$; Fig. 3*I*: $t_{(13)} = 0.4163$, $p = 0.6840$). Therefore, the mutations in the $NR1^{DATCreERT2}$ and $mGluR5^{KD\text{-}D1}$ strains caused a delay in decision time.

Reward magnitude discrimination and probability discounting

In the second experiment, we tested whether an ablation of glutamate receptors in the DA system influenced the discrimination of reward magnitude and discounting of the value of large outcomes caused by a decrease in the probability of large reward delivery. In this task, the animals were offered a choice between 10 or 20 μl of water (Fig. 5*A*). Each session began with 20 forced choice trials, during which the animals were familiarized with the choice outcomes, followed by 40 free choice trials. When both outcomes were deterministic and the animals were allowed to choose freely, the animals preferred the larger amount of water (5 d average ranged from 68.5% to 100%; mean, 92.6%; Fig. 5*B–D*). However, when the probability of receiving the larger reward gradually decreased, the preference for the large reward decreased accordingly, indicating that the animals perceived and adapted to the changes in the reward value (Fig. 5*B*: probability, $F_{(3,33)} = 39.53$; Fig. 5*C*: probability, $F_{(3,45)} = 109.92$; Fig. 5*D*: probability, $F_{(3,36)} = 109.92$; all $p < 0.0001$). Although no effects of the mutations were observed on probability discounting (Fig. 5*B*: genotype × probability, $F_{(3,33)} = 0.85$, $p = 0.4753$; genotype, $F_{(1,11)} = 0.0005$, $p = 0.9831$; Fig. 5*C*: genotype × probability, $F_{(3,45)} = 0.15$, $p = 0.9275$; genotype, $F_{(1,15)} = 0.67$, $p = 0.4250$; Fig. 5*D*: genotype × probability, $F_{(3,36)} = 1.77$, $p = 0.1706$; genotype, $F_{(1,12)} = 1.39$, $p = 0.2614$), the analysis of the reaction times revealed a large increase in the latency to choose during both the forced choice and free choice trials in the $NR1^{DATCreERT2}$ and $mGluR5^{KD\text{-}D1}$ mice (Fig. 6*A*, forced choice: genotype × probability, $F_{(3,33)} = 3.11$, $p = 0.0396$; genotype, $F_{(1,11)} = 67.02$, $p < 0.0001$; probability, $F_{(3,33)} = 0.97$, $p = 0.4193$; free choice: genotype × probability, $F_{(3,33)} = 1.81$, $p = 0.1642$; genotype, $F_{(1,11)} = 42.73$, $p < 0.001$; probability, $F_{(3,33)} = 0.66$, $p = 0.5816$; Fig. 6*B*, forced choice: genotype × probability, $F_{(3,45)} = 1.42$, $p = 0.2486$; genotype, $F_{(1,15)} = 21.96$, $p = 0.0003$; probability, $F_{(3,45)} = 5.40$, $p = 0.0029$; free choice: genotype × probability, $F_{(3,45)} = 0.10$, $p = 0.9605$; genotype, $F_{(1,15)} = 9.14$, $p = 0.0085$; probability, $F_{(3,45)} =$

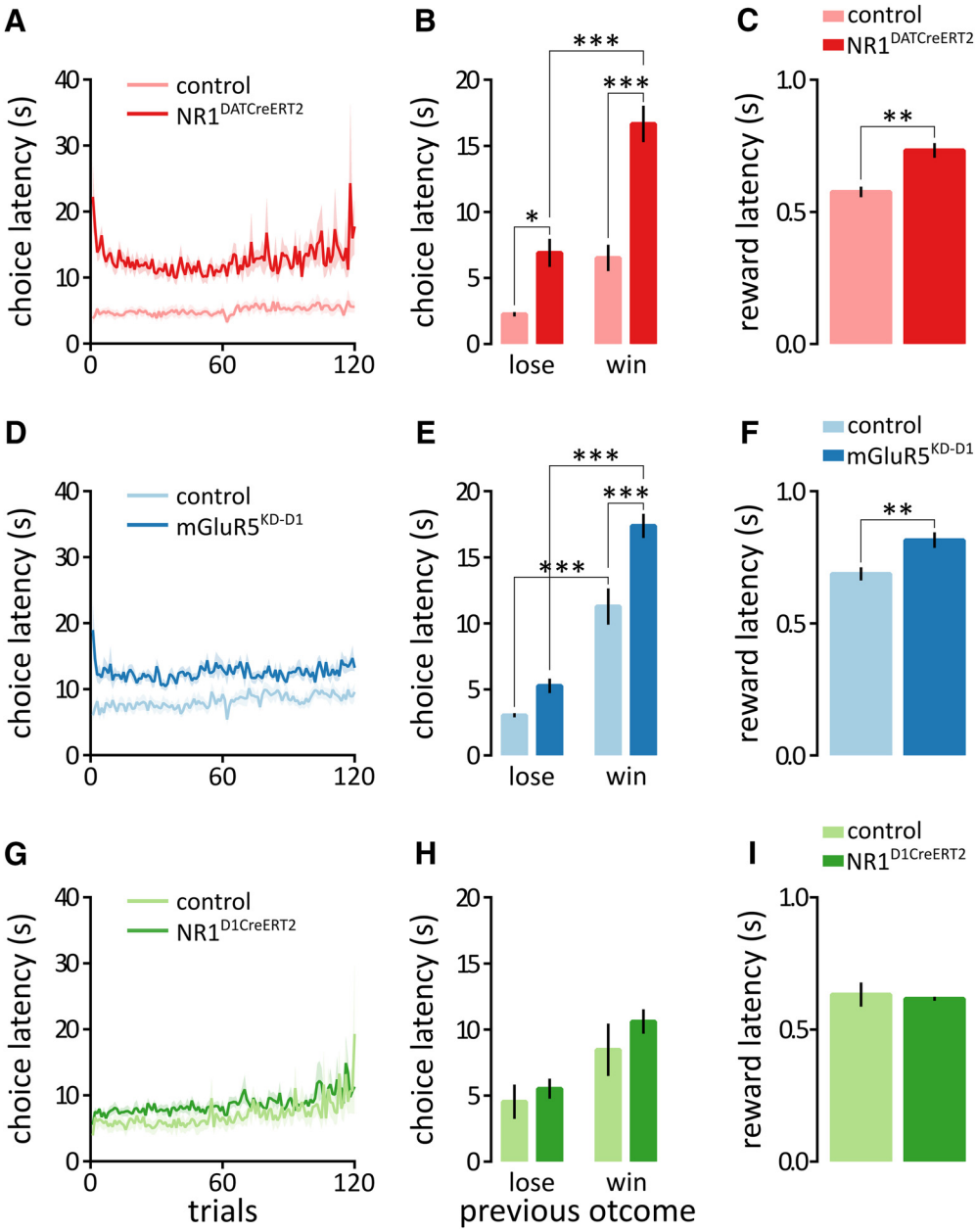


Figure 4. Reaction times in the probabilistic reinforcement learning task. *A–I*, Graphs show the reaction times observed in the NR1^{DATCreERT2} (mutant, $n = 6$; control, $n = 8$; *A–C*), mGluR5^{KD-D1} (mutant, $n = 8$; control, $n = 9$; *D–F*), and NR1^{D1CreERT2} (mutant, $n = 6$; control, $n = 9$; *G–I*) strains. *A*, *D*, and *G* show the time elapsed from the trial onset to the choice port entry. *B*, *E*, and *H* show the time from the new trial onset to the choice port entry following previously unrewarded (lose) or rewarded (win) trials. *C*, *F*, and *I* summarize the time from the reward delivery to the reward port entry. Values represent the mean choice latency (all sessions combined) ± SEM. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ (Bonferroni-corrected *t* test or *t* test).

6.11, $p = 0.0014$). This outcome was not observed in the case of the NR1^{D1CreERT2} mice (Fig. 6*C*, forced choice: genotype × probability, $F_{(3,36)} = 0.71$, $p = 0.5533$; genotype, $F_{(1,12)} = 0.53$, $p = 0.4815$; probability, $F_{(3,36)} = 8.94$, $p = 0.0001$; free choice: genotype × probability, $F_{(3,36)} = 2.61$, $p = 0.0665$; genotype, $F_{(1,12)} = 0.88$, $p = 0.3673$; probability, $F_{(3,36)} = 9.09$, $p = 0.0001$).

These results confirmed that while none of the mutations appreciably affected the magnitude discrimination or probability discounting, the animals from the NR1^{DATCreERT2} and mGluR5^{KD-D1} strains were considerably slower in performing choices.

Discussion

The mutations in the NR1^{DATCreERT2} and mGluR5^{KD-D1} strains had three effects on the choice behavior. First, the performance in the probabilistic reinforcement learning task was impaired, leading to fewer choices of the alternative with the higher reward probability. This effect was transient in the NR1^{DATCreERT2} strain, and the mutant mice

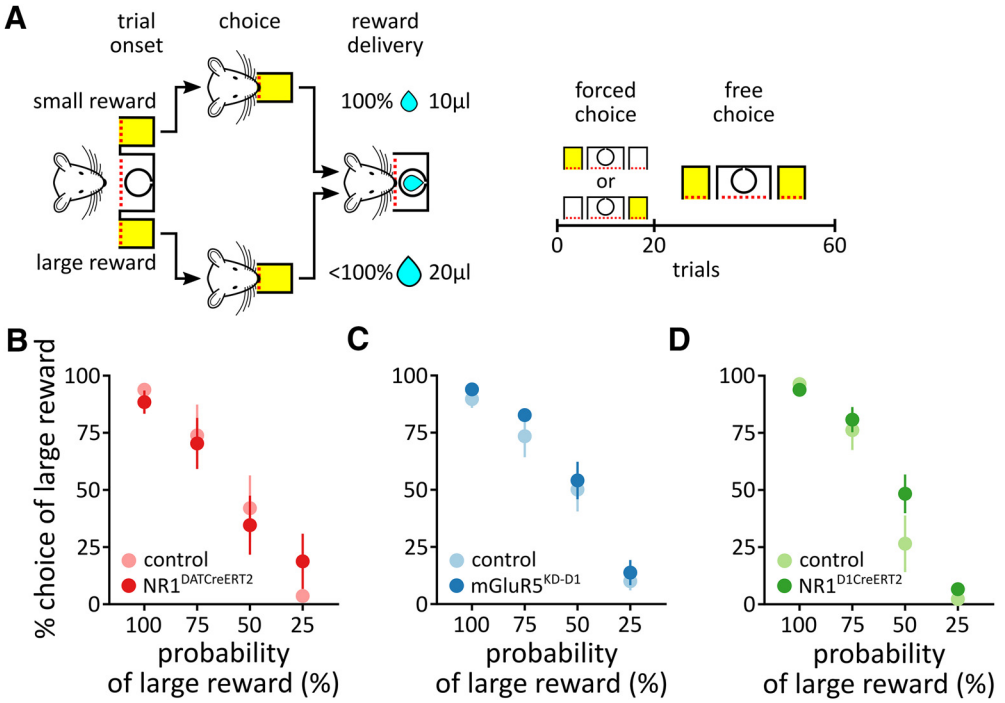


Figure 5. The probability-discounting task. *A*, Schematic representation of the probability-discounting task. One nose-poke port was associated with the delivery of small certain rewards, while the other nose-poke port was associated with the delivery of large uncertain rewards. Each session consisted of 20 forced trials during which only one port was active, followed by 40 free choice trials during which both ports were active. *B–D*, The graphs show the frequency of choosing the larger reward as a function of its probability in the NR1^{DATCreERT2} (mutant, $n = 6$; control, $n = 7$; *B*), mGluR5$^{KD\text{-}D1}$ (mutant, $n = 8$; control, $n = 9$; *C*), and NR1^{D1CreERT2} (mutant, $n = 5$; control, $n = 9$; *D*) strains. Data are represented as the mean ± SEM.

eventually reached the same performance as the controls, whereas the mGluR5$^{KD\text{-}D1}$ animals showed a generally lower preference for the higher value option. Second, the NR1^{DATCreERT2} and mGluR5$^{KD\text{-}D1}$ mice were less likely to repeat the previously rewarded choice. In accordance with this, computational modeling suggested that the behavior of both of these mutant groups was to a smaller extent influenced by acquired associations in comparison to controls (i.e., making more exploratory/random choices compared with controls). Finally, the third mutation effect in the NR1^{DATCreERT2} and mGluR5$^{KD\text{-}D1}$ strains was an increase in the delay to make a choice. In contrast, there were no appreciable changes in the behavior of the NR1^{D1CreERT2} mice.

Earlier studies have shown that the inactivation of functional NMDA receptors in DA neurons impaired burst firing and attenuated phasic DA release in the striatum (Zweifel et al., 2009; Parker et al., 2010; Wang et al., 2011). Consistent with this finding, we recently reported that the induction of the mutation in the NR1^{DATCreERT2} mice causes a complete loss of NMDA receptor-dependent bursting of midbrain DA neurons (Jastrzębska et al., 2016). Considering the role of DA neuron burst firing in reward prediction error coding (Schultz et al., 1997; Glimcher, 2011), the observed effects of the mutation are to an extent unexpected, as no significant changes in learning rates were observed. Still, we note that the reduced win-stay probability is actually similar to the effect reported in the case of optogenetic studies, where the inhibition of DA neurons imitating negative reward prediction error reduced the likelihood of returning to the previously rewarded alternative (Hamid et al., 2016; Parker et al., 2016). Moreover, the study by Pessiglione et al. (2006) offers a possible explanation for why the reduced bursting of DA neurons might have led to less deterministic behavior rather than a reduced learning rate. In that study, the effects of a drug-reducing DA function on learning in an analogous task was studied in humans inside an fMRI scanner. The authors developed a computational model that captured both behavioral data and blood oxygenation level-dependent responses in striatum, which are known to correlate with reward prediction error. According to this model, the drug had an effect of reducing the value of reward parameter r_t on trials where the reward is obtained (see Eqs. 1–4). Reducing r_t has exactly the same effect on model behavior as reducing inverse temperature β (identified in our study for NR1^{DATCreERT2} and mGluR5$^{KD\text{-}D1}$ mice) for the following reason. Reducing r_t decreases the value to which the estimators $V_{c,t}$ converge, because they approach the expected value of the reward. If both $V_{1,t}$ and $V_{2,t}$ are reduced by the same constant, this constant can be taken outside the bracket in the softmax Equation 5 and incorporated into β giving a lower effective value of β. Computational models with reduced r_t and β predict exactly the same behavior, and therefore cannot be distinguished on the basis of our data. Pessiglione et al. (2006) had additional neurophysiological data, indicating the value of reward prediction on

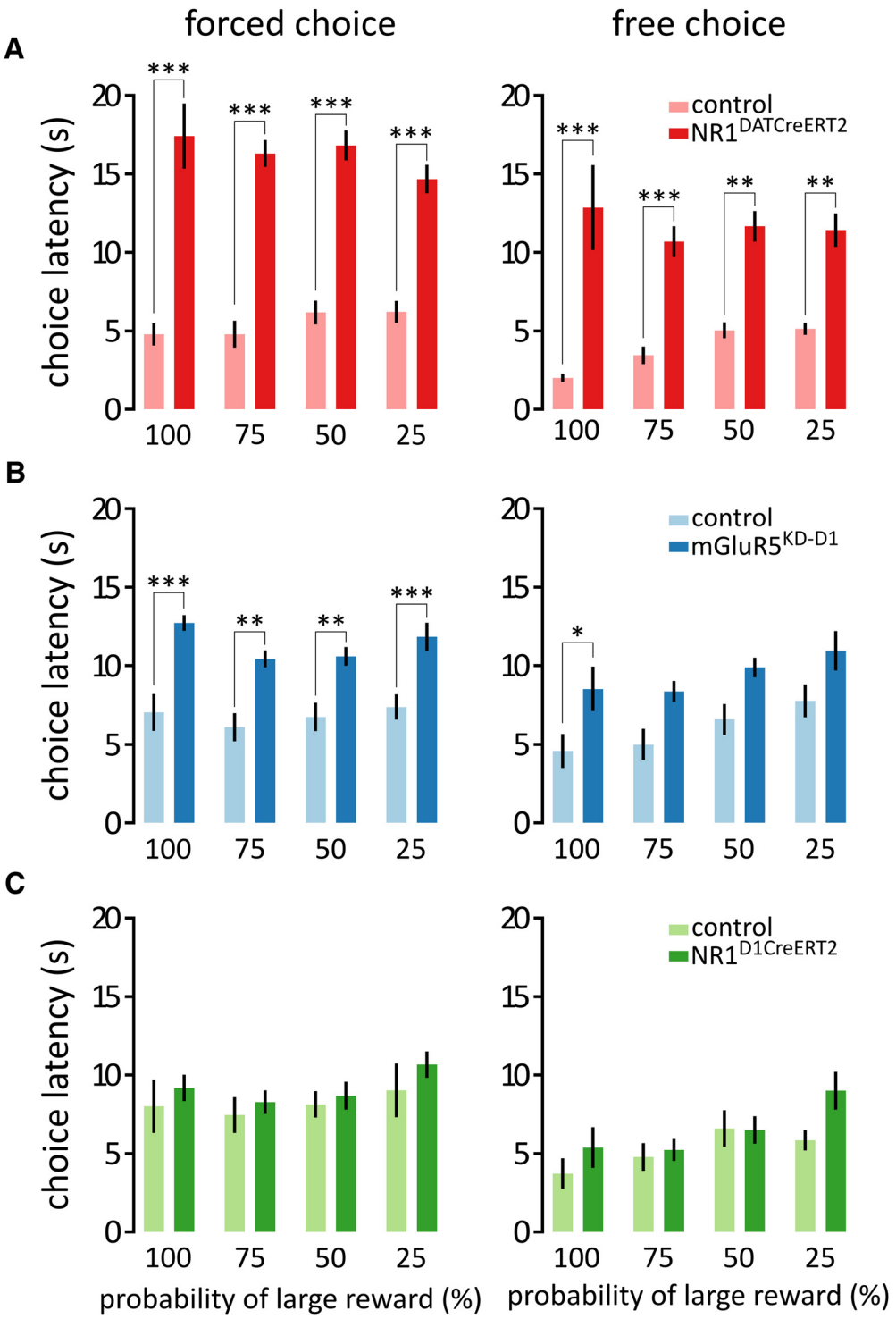


Figure 6. Reaction times in the probability-discounting task. *A–C*, Time elapsed from the trial onset to the choice port entry during the forced choice (left) and free choice (right) trials in the $NR1^{DATCreERT2}$ (mutant, $n = 6$; control, $n = 7$; *A*), $mGluR5^{KD\text{-}D1}$ (mutant, $n = 8$; control, $n = 9$; *B*), and $NR1^{D1CreERT2}$ (mutant, $n = 5$; control, $n = 9$; *C*) strains. Bars represent the mean choice latency ± SEM. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ (Bonferroni-corrected *t* test).

each trial, which allowed them to distinguish between these models. Thus, in summary, the less deterministic behavior of $NR1^{DATCreERT2}$ mice in our study might have resulted from impaired encoding of reward prediction error that led to reduced estimates of expected reward.

Nevertheless, we note that while the impairment caused by the mutation is clearly significant, it was arguably mild, and the $NR1^{DATCreERT2}$ mice eventually reached the same performance as that observed in the control animals. This is in agreement with the observation that

after extended training, performance levels in mice with constitutive mutations are similar to those found in control animals (Zweifel et al., 2009; James et al., 2015). Furthermore, in addition to its role in signaling reward prediction errors, phasic DA encodes expected reward value and contributes to risk-based decision-making (Fiorillo et al., 2003; Tobler et al., 2005; Sugam et al., 2012). This hypothesis is supported by observations in which the pharmacological blockade of DA receptors or the attenuation of phasic activity in DA neurons biases choices away from larger but probabilistic rewards (St Onge and Floresco, 2009; St Onge et al., 2011; Stopper et al., 2013, 2014). However, we found no effect of the loss of NMDA receptors on probability discounting, suggesting that NMDA receptors in DA neurons are not required for assessing the reward value when choosing between deterministic and probabilistic outcomes.

The inactivation of mGluR5 receptors in D_1-expressing neurons decreased the frequency of choosing the alternative with a higher reward probability. Thus, the mGluR5$^{KD\text{-}D1}$ mice made more random choices. Simultaneously, the NR1^{D1CreERT2} mice showed a normal performance. This result may be due to differences in the efficiency of the mutations in the dorsal part of the striatum. We have previously reported that a mutation in D_1CreERT2-derived strains is efficient in the nucleus accumbens and ventral striatum but is less extensive in the dorsal parts of the striatum (Rodriguez Parkitna et al., 2010; Sikora et al., 2016), whereas, in the mGluR5$^{KD\text{-}D1}$ strain, the mutation is efficient in both regions (Novak et al., 2010; Rodriguez Parkitna et al., 2013). The ventral components of the striatum are involved in stimulus-outcome learning, but the dorsal striatum plays a key role in learning about actions and their consequences (Balleine et al., 2007; Yin et al., 2008). A dissociable role of the ventral and dorsal striatal regions in choice behavior was also recently reported by Parker et al. (2016). These authors showed that DA terminals in the ventral striatum responded preferentially to reward consumption and reward-predicting cues, whereas terminals in the dorsal striatum responded more strongly to choices. Accordingly, optogenetic studies have demonstrated that the stimulation of D_1 neurons in the dorsal striatum mimic changes in action values and bias choice behavior during decision-making (Tai et al., 2012). Therefore, we speculate that when glutamate receptor-dependent plasticity is disrupted at corticostriatal synapses in the dorsal, rather than the ventral striatum, an increased randomness in action selection occurs.

The strongest effect observed in our study was the increased delay in performing a choice in the NR1^{DATCreERT2} and mGluR5$^{KD\text{-}D1}$ mice. This effect is consistent with a reported increase in the latency to choose in the appetitive T-maze task in NR1^{DATCre} mice (Zweifel et al., 2009) and the effect of optogenetic stimulation of DA neurons on the delay to engage in reward-seeking behavior (Hamid et al., 2016). Notably, our procedure imposed no limit on the trial length, while a 10 s limit was often used previously (Stopper et al., 2014; Parker et al., 2016). If a limit had been imposed, we would have likely observed a large number of omitted trials. Thus, a decision time limit could likely exacerbate the phenotypes observed in the probabilistic reinforcement learning task. It should also be noted that the mutations affected the time to collect the reward. However, only a slight increase in the reward latency was observed. The influence of the mutations on locomotor activity in this case seems to be rather unlikely. First, it was previously reported that a mutation in NR1^{DATCreERT2} mice had no effect on locomotor activity in the home cage or open field arena (Engblom et al., 2008), and only a mild reduction of activity in the novel environment was observed in mGluR5$^{KD\text{-}D1}$ mice, with no change in the distance traveled in familiar environment (Rodriguez Parkitna et al., 2013). Second, based on the performance in the rotarod test, there is no evidence of motor impairment in NR1^{DATCreERT2} mice (Jastrzębska et al., 2016). We thus believe that an increase in choice latency is a result of an internal decision (or motivational) process, rather than a result of impaired motor performance. This interpretation is in line with observations showing that perturbations in mesolimbic DA signaling result in decreased motivation to engage in reward-seeking behavior, which is expressed as an increase in latency to initiate instrumental phase of reward-directed behavior (Nicola, 2010; Salamone and Correa, 2012).

In conclusion, we find that the loss of NMDA receptors in DA neurons and mGluR5 receptors in D_1-expressing neurons affects the speed of the decision process and increases the number of exploratory choices. Nevertheless, mutant mice did improve their performance in the probabilistic reinforcement learning task and showed normal probability discounting. Overall, this indicates that reward-driven learning does occur in the absence of key receptors implicated in the plasticity of the reward system of the brain, but the decision-making process slows and loses efficiency.

References

Ahn W-Y, Busemeyer JR, Wagenmakers E-J, Stout JC (2008) Comparison of decision learning models using the generalization criterion method. Cogn Sci 32:1376–1402. CrossRef Medline

Ahn W-Y, Krawitz A, Kim W, Busmeyer JR, Brown JW (2011) A model-based fMRI analysis with hierarchical bayesian parameter estimation. J Neurosci Psychol Econ 4:95–110. CrossRef Medline

Ahn W-Y, Vasilev G, Lee S-H, Busemeyer JR, Kruschke JK, Bechara A, Vassileva J (2014) Decision-making in stimulant and opiate addicts in protracted abstinence: evidence from computational modeling with pure users. Front Psychol 5:849. CrossRef Medline

Ahn W-Y, Haines N, Zhang L (2017) Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. Comput Psychiatry 1:24–57. CrossRef Medline

Balleine BW, Delgado MR, Hikosaka O (2007) The role of the dorsal striatum in reward and decision-making. J Neurosci 27:8161–8165. CrossRef Medline

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141. CrossRef Medline

Berridge KC, Robinson TE (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?. Brain Res Brain Res Rev 28:309–369. Medline

Beutler LR, Eldred KC, Quintana A, Keene CD, Rose SE, Postupna N, Montine TJ, Palmiter RD (2011) Severely impaired learning and altered neuronal morphology in mice lacking NMDA receptors in medium spiny neurons. PloS One 6:e28168. CrossRef Medline

Calabresi P, Picconi B, Tozzi A, Di Filippo M (2007) Dopamine-mediated regulation of corticostriatal synaptic plasticity. Trends Neurosci 30:211–219. CrossRef Medline

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: a probabilistic programming language. J Stat Softw 76:1–32. CrossRef

Chergui K, Charléty PJ, Akaoka H, Saunier CF, Brunet JL, Buda M, Svensson TH, Chouvet G (1993) Tonic activation of NMDA receptors causes spontaneous burst discharge of rat midbrain dopamine neurons in vivo. Eur J Neurosci 5:137–144. Medline

den Ouden HEM, Daw ND, Fernandez G, Elshout JA, Rijpkema M, Hoogman M, Franke B, Cools R (2013) Dissociable effects of dopamine and serotonin on reversal learning. Neuron 80:1090–1100. CrossRef Medline

Engblom D, Bilbao A, Sanchis-Segura C, Dahan L, Perreau-Lenz S, Balland B, Rodriguez Parkitna J, Luján R, Halbout B, Mameli M, Parlato R, Sprengel R, Lüscher C, Schütz G, Spanagel R (2008) Glutamate receptors on dopamine neurons control the persistence of cocaine seeking. Neuron 59:497–508. CrossRef

Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299:1898–1902. CrossRef Medline

Floresco SB, Magyar O (2006) Mesocortical dopamine modulation of executive functions: beyond working memory. Psychopharmacology (Berl) 188:567–585. CrossRef Medline

Gläscher J, Hampton AN, O'Doherty JP (2009) Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cereb Cortex 19:483–495. CrossRef

Glimcher PW (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci U S A 108:15647–15654. CrossRef

Grogan JP, Tsivos D, Smith L, Knight BE, Bogacz R, Whone A, Coulthard EJ (2017) Effects of dopamine on reinforcement learning and consolidation in Parkinson's disease. Elife 6:e26801. CrossRef

Hamid AA, Pettibone JR, Mabrouk OS, Hetrick VL, Schmidt R, Vander Weele CM, Kennedy RT, Aragona BJ, Berke JD (2016) Mesolimbic dopamine signals the value of work. Nat Neurosci 19:117–126. CrossRef Medline

Harnett MT, Bernier BE, Ahn K-C, Morikawa H (2009) Burst-timing-dependent plasticity of NMDA receptor-mediated transmission in midbrain dopamine neurons. Neuron 62:826–838. CrossRef Medline

Hart AS, Rutledge RB, Glimcher PW, Phillips PEM (2014) Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J Neurosci 34:698–704. CrossRef

James AS, Pennington ZT, Tran P, Jentsch JD (2015) Compromised NMDA/glutamate receptor expression in dopaminergic neurons impairs instrumental learning, but not pavlovian goal tracking or sign tracking. eNeuro 2:0040-14.2015. CrossRef

Jastrzębska K, Walczak M, Cieślak PE, Szumiec Ł, Turbasa M, Engblom D, Błasiak T, Rodriguez Parkitna J (2016) Loss of NMDA receptors in dopamine neurons leads to the development of affective disorder-like symptoms in mice. Sci Rep 6:37171. CrossRef Medline

Kruschke J (2014) Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. San Diego, CA: Academic.

Lee D, Seo H, Jung MW (2012) Neural basis of reinforcement learning and decision making. Annu Rev Neurosci 35:287–308. CrossRef Medline

Lee MD (2011) How cognitive modeling can benefit from hierarchical Bayesian models. J Math Psychol 55:1–7. CrossRef

Nicola SM (2010) The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. J Neurosci 30:16585–16600. CrossRef Medline

Novak M, Halbout B, O'Connor EC, Rodriguez Parkitna J, Su T, Chai M, Crombag HS, Bilbao A, Spanagel R, Stephens DN, Schütz G, Engblom D (2010) Incentive learning underlying cocaine-seeking requires mGluR5 receptors located on dopamine D_1 receptor-expressing neurons. J Neurosci 30:11973–11982. CrossRef

Overton P, Clark D (1992) Iontophoretically administered drugs acting at the N-methyl-D-aspartate receptor modulate burst firing in A9 dopamine neurons in the rat. Synapse 10:131–140. CrossRef

Parker JG, Zweifel LS, Clark JJ, Evans SB, Phillips PEM, Palmiter RD (2010) Absence of NMDA receptors in dopamine neurons attenuates dopamine release but not conditioned approach during Pavlovian conditioning. Proc Natl Acad Sci U S A 107:13491–13496. CrossRef

Parker JG, Beutler LR, Palmiter RD (2011) The contribution of NMDA receptor signaling in the cortico-basal ganglia reward network to appetitive Pavlovian learning. J Neurosci 31:11362–11369. CrossRef

Parker NF, Cameron CM, Taliaferro JP, Lee J, Choi JY, Davidson TJ, Daw ND, Witten IB (2016) Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat Neurosci 19:845–854. CrossRef

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442:1042–1045. CrossRef Medline

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II: Current Research and Theory, Vol. 2, pp 64–99.

Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. Nature 413:67–70. CrossRef Medline

Rodriguez Parkitna J, Bilbao A, Rieker C, Engblom D, Piechota M, Nordheim A, Spanagel R, Schütz G (2010) Loss of the serum response factor in the dopamine system leads to hyperactivity. FASEB J 24:2427–2435. CrossRef

Rodriguez Parkitna J, Sikora M, Gołda S, Gołembiowska K, Bystrowska B, Engblom D, Bilbao A, Przewlocki R (2013) Novelty-seeking behaviors and the escalation of alcohol drinking after abstinence in mice are controlled by metabotropic glutamate receptor 5 on neurons expressing dopamine d1 receptors. Biol Psychiatry 73:263–270. CrossRef

Salamone JD, Correa M (2012) The mysterious motivational functions of mesolimbic dopamine. Neuron 76:470–485. CrossRef Medline

Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337–1340. CrossRef Medline

Schultz W (2015) Neuronal reward and decision signals: from theories to data. Physiol Rev 95:853–951. CrossRef Medline

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599. Medline

Shen W, Flajolet M, Greengard P, Surmeier DJ (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321:848–851. CrossRef Medline

Shiffrin RM, Lee MD, Kim W, Wagenmakers E-J (2008) A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. Cogn Sci 32:1248–1284. CrossRef Medline

Shiner T, Seymour B, Wunderlich K, Hill C, Bhatia KP, Dayan P, Dolan RJ (2012) Dopamine and performance in a reinforcement learning task: evidence from Parkinson's disease. Brain 135:1871–1883. CrossRef Medline

Sikora M, Tokarski K, Bobula B, Zajdel J, Jastrzębska K, Cieślak PE, Zygmunt M, Sowa J, Smutek M, Kamińska K, Gołembiowska K, Engblom D, Hess G, Przewlocki R, Rodriguez Parkitna J (2016) NMDA receptors on dopaminoceptive neurons are essential for drug-induced conditioned place preference. eNeuro 3:ENEURO.0084-15.2016.

Sombers LA, Beyene M, Carelli RM, Wightman RM (2009) Synaptic overflow of dopamine in the nucleus accumbens arises from neu-

ronal activity in the ventral tegmental area. J Neurosci 29:1735–1742. CrossRef Medline

St Onge JR, Floresco SB (2009) Dopaminergic modulation of risk-based decision making. Neuropsychopharmacol 34:681–697. CrossRef

St Onge JR, Abhari H, Floresco SB (2011) Dissociable contributions by prefrontal D_1 and D_2 receptors to risk-based decision making. J Neurosci 31:8625–8633. CrossRef Medline

Steingroever H, Wetzels R, Wagenmakers E (2014) Absolute performance of reinforcement-learning models for the Iowa Gambling Task. Decision 1: 161–183. CrossRef

Stopper CM, Khayambashi S, Floresco SB (2013) Receptor-specific modulation of risk-based decision making by nucleus accumbens dopamine. Neuropsychopharmacology 38:715–728. CrossRef

Stopper CM, Tse MTL, Montes DR, Wiedman CR, Floresco SB (2014) Overriding phasic dopamine signals redirects action selection during risk/reward decision making. Neuron 84:177–189. CrossRef Medline

Stuber GD, Klanker M, de Ridder B, Bowers MS, Joosten RN, Feenstra MG, Bonci A (2008) Reward-predictive cues enhance excitatory synaptic strength onto midbrain dopamine neurons. Science 321:1690–1692. CrossRef Medline

Sugam JA, Day JJ, Wightman RM, Carelli RM (2012) Phasic nucleus accumbens dopamine encodes risk-based decision-making behavior. Biol Psychiatry 71:199–205. CrossRef Medline

Surmeier DJ, Plotkin J, Shen W (2009) Dopamine and synaptic plasticity in dorsal striatal circuits controlling action selection. Curr Opin Neurobiol 19:621–628. CrossRef

Tai L-H, Lee AM, Benavidez N, Bonci A, Wilbrecht L (2012) Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat Neurosci 15:1281–1289. CrossRef Medline

Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307:1642–1645. CrossRef Medline

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27:1413–1432. CrossRef

Wang LP, Li F, Wang D, Xie K, Wang D, Shen X, Tsien JZ (2011) NMDA receptors in dopaminergic neurons are crucial for habit learning. Neuron 72:1055–1066. CrossRef Medline

Wickham R, Solecki W, Rathbun L, McIntosh JM, Addy NA (2013) Ventral tegmental area $\alpha 6\beta 2$ nicotinic acetylcholine receptors modulate phasic dopamine release in the nucleus accumbens core. Psychopharmacology (Berl) 229:73–82. CrossRef Medline

Wise RA (2004) Dopamine, learning and motivation. Nat Rev Neurosci 5:483–494. CrossRef Medline

Yagishita S, Hayashi-Takagi A, Ellis-Davies GCR, Urakubo H, Ishii S, Kasai H (2014) A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science 345:1616–1620. CrossRef

Yin HH, Ostlund SB, Balleine BW (2008) Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. Eur J Neurosci 28:1437–1448. CrossRef Medline

Zweifel LS, Parker JG, Lobb CJ, Rainwater A, Wall VZ, Fadok JP, Darvas M, Kim MJ, Mizumori SJY, Paladini CA, Phillips PEM, Palmiter RD (2009) Disruption of NMDAR-dependent burst firing by dopamine neurons provides selective assessment of phasic dopamine-dependent behavior. Proc Natl Acad Sci U S A 106:7281–7288. CrossRef