Psychoneuroendocrinology 100 (2019) 27–31

Contents lists available at ScienceDirect

Psychoneuroendocrinology

journal homepage: www.elsevier.com/locate/psyneuen

Short Communication

Reinforcement learning across the rat estrous cycle

Jeroen P.H. Verharen[a,b], Jiska Kentrop[a], Louk J.M.J. Vanderschuren[b,*,1], Roger A.H. Adan[a,*,1]

[a] Brain Center Rudolf Magnus, Department of Translational Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands
[b] Department of Animals in Science and Society, Division of Behavioural Neuroscience, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Learning
Motivation
Estrous cycle
Hormones
Rats
Reinforcement earning

ABSTRACT

Reinforcement learning, the process by which an organism flexibly adapts behavior in response to reward and punishment, is vital for the proper execution of everyday behaviors, and its dysfunction has been implicated in a wide variety of mental disorders. Here, we use computational trial-by-trial analysis of data of female rats performing a probabilistic reward learning task and demonstrate that core computational processes underlying value-based decision making fluctuate across the estrous cycle, providing a neuroendocrine substrate by which gonadal hormones may influence adaptive behavior.

1. Introduction

Reinforcement learning is an essential mechanism for organisms to adapt to a dynamic environment, by allowing flexible alterations in behavior in response to positive and negative feedback, for example during foraging and social encounters (Sutton and Barto, 1998). As such, deficits in reinforcement learning have been implicated in several psychiatric conditions, including addiction and schizophrenia (Maia and Frank, 2011). Given the large gender differences in the prevalence of mental disorders, and the existence of cyclic changes in the severity of schizophrenia and sensitivity to drugs in women (Hendrick et al., 1996), we sought to determine how the estrous cycle of females affects the computational processes that underlie reinforcement learning. To this aim, we tested a cohort of female rats on a probabilistic reversal learning paradigm (Bari et al., 2010; Verharen et al., 2018), used computational modeling to extract the subcomponents of value-based decision making, and assessed how these components were affected by the estrous cycle.

2. Methods

2.1. Animals

Female, nulliparous Long-Evans rats (bred in-house; background Rj:Orl, Janvier labs, France; $n = 30$) weighing 180–220 g were used for the experiment. Animals were tested for 10 consecutive days, to ensure that we had at least one measurement of every cycle stage per animal. Eventually, 5 animals had to be excluded because the cycle could not reliably be estimated or not all stages of the cycle were captured due to unreliable vaginal smears, leaving a final group of $n = 25$. Animals were socially housed in groups of 2–4 and kept on a reversed day/night cycle (lights on at 8 A.M.), and behavioral experiments took place between 9 A.M. and 1 P.M.. During the training phase of the experiment, animals were kept on a food restriction regimen of 5 g chow per 100 g body weight, and during the 10 experimental days the animals were food restricted for 16 h prior to the behavioral task. For the male group of animals ($n = 18$), that is included for comparison, Long-Evans rats (bred in-house; background Rj:Orl, Janvier labs, France) of roughly the same age, weighing 310–390 g, were used. Animals had *ad libitum* access to water, except during behavioral experiments. The experiments were carried out in accordance with Dutch legislation (Wet op de Dierproeven, 2014), European Union guidelines (2010/63/EU), and approved by the Animal Welfare Body of Utrecht University and the Dutch Central Animal Testing Committee.

2.2. Behavioral task

The probabilistic reversal learning task (Fig. 1a) took place in operant conditioning chambers (Med Associates Inc., USA) equipped with a food receptacle (with infra-red entry detection) flanked by two retractable levers and two cue lights, a house light and an auditory tone generator. One lever was randomly assigned as the high-probability lever, responding on which was reinforced (i.e., delivery of a sucrose pellet) with an 80% probability and not reinforced (i.e., a time-out) with a 20% probability. The other lever was assigned as the low-probability lever, responding on which had a 20% chance of being

* Corresponding authors.
E-mail addresses: l.j.m.j.vanderschuren@uu.nl (L.J.M.J. Vanderschuren), r.a.h.adan@umcutrecht.nl (R.A.H. Adan).
[1] These authors contributed equally.

https://doi.org/10.1016/j.psyneuen.2018.09.016
Received 25 May 2018; Received in revised form 24 July 2018; Accepted 12 September 2018

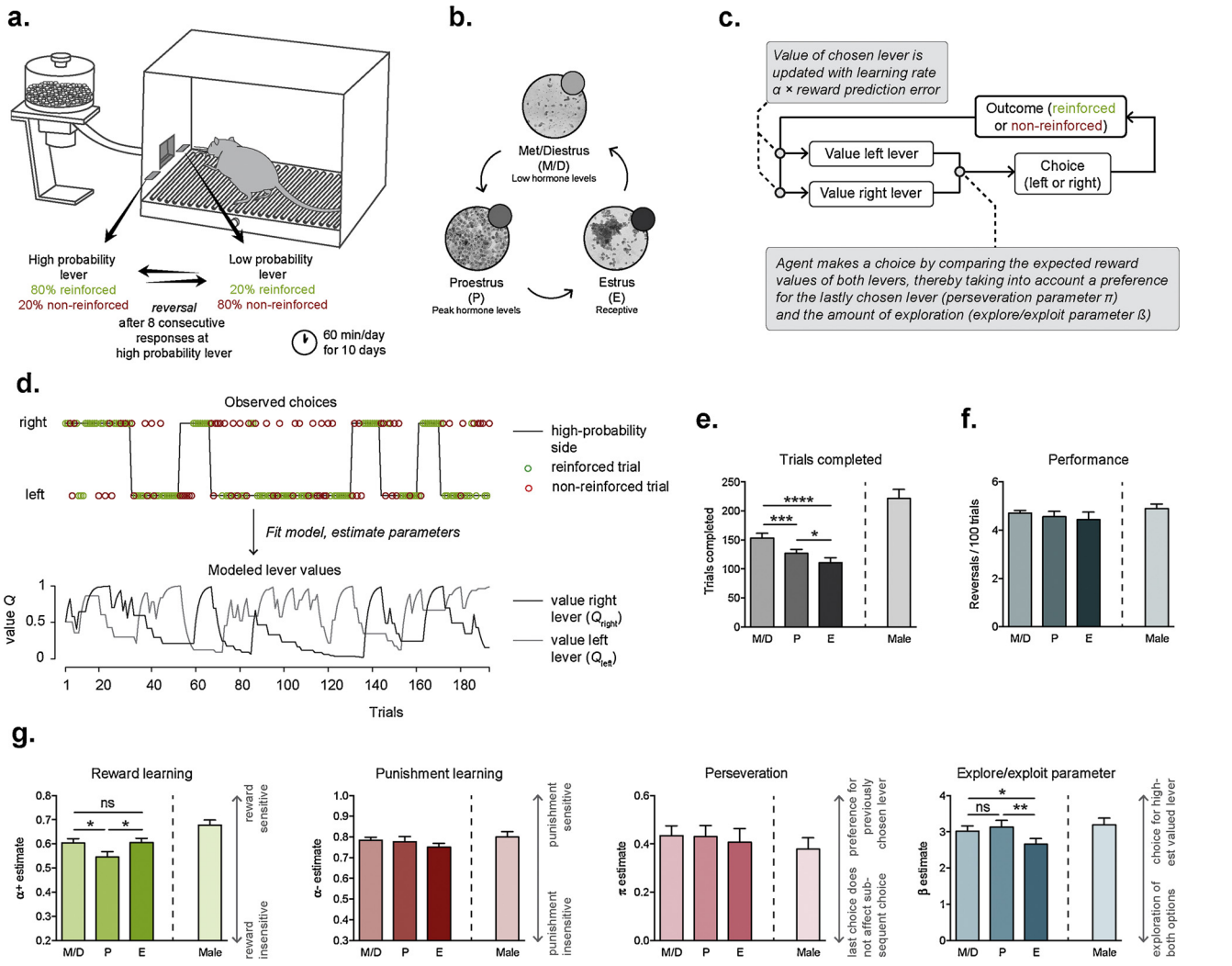

Fig. 1. a. Probabilistic reversal learning setup. Hungry female animals could respond on two levers, one of which delivered sucrose reward with a high probability (80%, high-probability lever), and the other lever with a low probability (20%, low-probability lever). Every time the animal made eight consecutive responses on the high-probability lever, a reversal in reinforcement contingencies occurred, so that the previously low-probability lever became the high-probability lever, and vice versa. In this way, animals had to track the outcome of responding on each of the two levers over a series of trials and based hereon make a choice between them. b. Example cytological images of samples from vaginal smears during the three stages of the estrous cycle. c. Computational model. d. Trial-to-trial data was fit to the computational model, and best-fit parameters were estimated. e. Total trials completed by the female animals ($n = 25$) in the 60-minute session was significantly affected by the estrous cycle (Repeated measures ANOVA, $F_{2,48} = 21.22$, $P < 0.0001$). Post-hoc tests: **** $P < 0.0001$, *** $P = 0.0002$, * $P = 0.0188$. Male data ($n = 18$) is shown for illustrative purposes; these data were not included in the statistical analyses. f. The total number of reversals was not affected by the cycle (ANOVA, $F_{2,48} = 0.48$, $P = 0.6209$). g. Best-fit computational model parameters per estrous cycle stage. Reward learning: ANOVA $F_{2,48} = 3.995$, $P = 0.0248$; post-hoc tests met/diestrus (M/D) vs proestrus (P), $P = 0.0198$, M/D vs estrus (E), $P = 0.9425$, P vs E, $P = 0.0166$. Punishment learning: ANOVA $F_{2,48} = 1.637$, $P = 0.2052$. Perseveration: ANOVA $F_{2,48} = 0.1349$, $P = 0.8741$. Explore/exploit: ANOVA $F_{2,48} = 5.201$, $P = 0.0090$; post-hoc tests M/D vs P, $P = 0.4444$, M/D vs E, $P = 0.0243$, P vs E, $P = 0.0033$. Male data is shown for illustrative purposes.

reinforced. Every single response on the high-probability and low-probability lever was reinforced with a 80% or 20% probability, respectively, irrespective of the outcome of the previous trials.

The session lasted for 60 min, and animals were constrained in the number of trials they could make only by the length of the session (maximum ~600 trials per session possible). A trial commenced by the illumination of the house light and the presentation of the two levers into the operant cage. After a lever press by the animal, the levers retracted and the house light was turned off. For reinforced trials, a 45 mg sucrose pellet (5TUL, TestDiet, USA) was delivered into the food port, and both cue lights that flanked the food receptacle were illuminated, and an auditory tone was played for 0.5 s. A new trial commenced directly when the animal entered the food port (detected by the infra-red movement detector); this was signaled to the animal by extinction of the cue lights, illumination of the house light and presentation of the two levers. On non-reinforced trials, no additional cues were presented, leaving the animals in the dark during a 10 s period.

Every time the animal made 8 consecutive responses on the high-probability lever, a reversal in reinforcement contingencies occurred, so that the high-probability and low-probability levers switched. This reversal was not signaled to the animal, so it had to infer this contingency switch from the outcomes of the trials.

The software automatically registered the responses and response times of the animals, as well as the outcome of the trial (reinforced or not), and the position of the high-probability lever.

2.3. Training

Animals first received lever press training, during which both levers were continuously presented, and a lever press was reinforced under a

fixed ratio-1 schedule of reinforcement. When all animals made more than 50 lever presses in a session, the group progressed to the next phase of lever press training, in which randomly the left lever, the right lever, or both levers were presented to the animals, and pressing either lever was reinforced under a fixed ratio-1 schedule. In this phase of training, levers retracted after a response, and animals were subjected to the same sequence of events as during a reinforced trial in the probabilistic reversal learning task. When all animals made at least 100 responses in a session during this phase, the group received 6 training sessions of the probabilistic reversal learning task, before the experimental phase began (both females and males received these 6 training sessions in the final stage).

2.4. Estrous cycle determination

To determine the circulating levels of female sex hormones throughout the estrous cycle, vaginal smears were obtained for all test days between 11 A.M. and 1 P.M., 1–2 h after each test. Vaginal smears were collected by inserting the head of a sterile plastic smear loop (1 μL; VWR, USA) and gently swabbing the vaginal wall. The collected cells were transferred to a drop of water on a glass microscope slide, air-dried and stained with 5% Giemsa (Sigma-Aldrich, The Netherlands) dissolved in water. Microscopic evaluation of the cells present in the vaginal smears was used to determine the phase of the estrous cycle (Cora et al., 2015; Goldman et al., 2007) (Fig. 1b). This was performed by a trained observer who was blind to smears from previous days and the behavioral data, and the following four parameters were estimated: the relative amount of cells present (on a scale from 1 to 5), and the percentage of nucleated cells, anucleated cells and leukocytes. Based on these four parameters and taking into account all 10 days, smears were assigned as proestrus, estrus or metestrus-diestrus. In brief; smears containing predominantly nucleated cells were assigned as proestrus, smears containing predominantly anucleated cells were assigned as estrus and smears containing leukocytes were assigned as metestrus-diestrus. Smears containing a combination of cells indicating a transition between phases were interpreted based on smears from neighboring days. Females that did not show a regular cycle over the course of 10 days were excluded from the analysis. If a single smear was unreliable for a given day, but smears of neighboring days showed a predictable pattern coherent with a regular estrous cycle, the phase of the missing day was estimated; if not, that particular day was not included in the analysis.

2.5. Reinforcement learning model

The trial-by-trial data of every individual session was fit to a reinforcement learning model, which was a modification of the classic Rescorla-Wagner model (Rescorla and Wagner, 1972), which assumes that the animals dynamically track the value of the outcome of responding on each of the two levers by incorporating positive (reward delivery) and negative (reward omission) feedback (Fig. 1c, d). When learning from feedback is high ($\alpha \rightarrow 1$), these lever values are strongly dependent on the outcome of the last trial, but when learning is low ($\alpha \rightarrow 0$), lever values are based on an extended history of trials (thus the impact of a single reward delivery or reward omission on lever value is small). The model further incorporates the animals' preference for the lastly chosen lever, independent of lever values, which is captured by perseveration parameter π. Moreover, it incorporates stochastic choice, to distinguish between deterministic choice of the highest valued lever ($\beta \rightarrow \infty$) and a more exploratory sampling approach ($\beta \rightarrow 0$). Random effects model selection indicated that this modified Rescorla-Wagner model was able to predict the highest amount of observed choices compared to a set of other reinforcement learning models that we tested, including the classic Rescorla-Wagner model (Rescorla and Wagner, 1972), a Pearce-Hall-Rescorla-Wagner hybrid model (Li et al., 2011), and a win-stay, lose-switch model (Posch, 1999) (Supplementary Table 1).

The expected reward values of both levers, Q_{left} and Q_{right}, ranged from 0 (pressing the lever is never reinforced) to 1 (pressing the lever is always reinforced). Both lever values were initiated at a value of 0.5, and the value of the chosen lever Q_{chosen} was updated after every trial t based on the outcome of that trial:

$$Q_{chosen,t} = \begin{cases} Q_{chosen,t-1} + \alpha^{+} \cdot \delta_{t-1} & \textit{for rewarded trials} \\ Q_{chosen,t-1} + \alpha^{-} \cdot \delta_{t-1} & \textit{for time-out trials} \end{cases}$$

Here, α^{+} is the reward learning rate (learning from positive feedback), and α^{-} is the punishment learning rate (learning from negative feedback), which range from 0 (no learning) to 1 (lever value completely determined by last outcome). $\delta_{\text{t-1}}$ represents the reward prediction error after the last trial *t-1*, so that:

$$\delta_{t-1} = \begin{cases} 1 - Q_{chosen,t-1} & \textit{for rewarded trials} \\ 0 - Q_{chosen,t-1} & \textit{for time-out trials} \end{cases}$$

Note that reward prediction error δ is negative for non-reinforced trials (outcome is lower than expected) and positive for reinforced trials (outcome is higher than expected). The value of the unchosen lever was not updated. Separate learning rates were used for learning from positive feedback (i.e., $\delta > 0$; rewarded trials) versus negative feedback (i.e., $\delta < 0$; time-out trials), so that changes in reward or punishment learning could be discerned.

At the start of each trial, lever values Q_{left} and Q_{right} were converted to action probabilities using a Softmax function, so that the probability of choosing the right lever $p_{\text{right,t}}$ at trial t was given by the function:

$$p_{right,t} = \frac{exp(\beta \cdot Q_{right,t} + \pi \cdot \phi_{right,t})}{exp(\beta \cdot Q_{left,t} + \pi \cdot \phi_{left,t}) + exp(\beta \cdot Q_{right,t} + \pi \cdot \phi_{right,t})}$$

Here, β is the inverse temperature of the Softmax function, which is a measure for the extent to which the animal consistently chooses the highest valued lever ($\beta \rightarrow \infty$) or that it chooses more randomly ($\beta \rightarrow 0$). Parameter π is a stickiness parameter, which adds a certain amount of the value of π to the value estimate of the lastly chosen lever. In this case, positive values of π indicate a preference for the lastly chosen lever, negative values of π indicate a preference for the lastly unchosen lever, and π approaching 0 indicates that the side of the lastly chosen lever does not affect the next lever choice. ϕ is a boolean that was attributed the value 1 if that lever was chosen in the last trial (thus an amount of the value of π will be added to the value function), and 0 if that lever was not chosen in the last trial.

To obtain reliable model parameter estimates on a population level, we used Bayesian hierarchical parameter estimation. In brief, we applied a prior distribution over the parameter values, and considered any new evidence from the animal's choice behavior to determine a posterior probability using Bayes' rule. These posterior probabilities were marginalized to get a point estimate of each session's best-fit parameter values. The used priors were: for α^{+} and α^{-} betapdf(1.5, 1.5); for π normpdf(0.5, 0.5); for β normpdf(2, 2).

All computational analyses were performed with Matlab R2014a (MathWorks Inc., USA).

2.6. Statistics

Statistical tests were performed in GraphPad Prism 6.0 (GraphPad Inc., USA). On all outcome parameters, a one-way repeated measures analysis of variance (one-way RM ANOVA) was performed, with estrous phase as a within-subjects repeated measures factor. This test was considered significant if $P < 0.05$, after which post-hoc Fisher's tests were performed. When data of more than one test per estrous phase was obtained (because data was collected from more than one cycle and/or animals were in a certain phase of the estrous cycle for more than one day), the outcome parameter values were averaged for these days. No statistical comparisons were made between males and females because

the two groups were not tested in parallel and therefore equal testing conditions could not be ensured. In all graphs: **** $P < 0.0001$, *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, ^{ns} not significant.

3. Results

We observed a significant effect of estrous cycle on the total number of trials that the animals made during a session (Fig. 1e). Animals that were in the estrus stage of the cycle made the lowest number of trials, and animals in the metestrus/diestrus stage the highest number of trials.

Performance in the task, measured as the total number of reversals that the animals achieved, revealed no significant differences between the three stages (Fig. 1f). However, the total number of reversals is a compound measure for performance in the task, that does not necessarily inform about the underlying component processes. To gain insight into whether these underlying processes were modulated by the cycle, we fit the trial-by-trial data in the session to a computational reinforcement learning model (Gershman, 2016), and used Bayesian hierarchical estimation (Daw, 2009) to determine the parameter values that best described the behavior of the animals (Fig. 1d). After estimating the value of the four model parameters for each session, and comparing these between the different stages of the cycle (Fig. 1g), we observed a significant decrease in reward learning parameter $\alpha+$ during the proestrus stage, indicative of a lower impact of positive feedback (i.e., reinforcement) on behavior. We further found that the estimate of explore/exploit parameter β was significantly reduced during the estrus stage. No significant changes were observed on the value estimates of punishment learning parameter α- and perseveration parameter π. We replicated these findings by fitting the data to a less complex model that only includes $\alpha+$, α- and β as free parameters (Supplementary Fig. 1). Overall, the value estimates of the parameters in female animals were roughly similar to those observed in males (Fig. 1g), except that male animals made more trials in the task (Fig. 1e).

4. Discussion

Our computational analyses reveal distinct changes in the processes underlying value-based decision making across the rat estrous cycle. The observed decrease in reward learning parameter $\alpha+$ during the proestrus stage is indicative of a lower impact of positive feedback (i.e., reinforcement) on behavior. This stage of the cycle is characterized by peak levels of the sex hormones progesterone and estradiol, and thus suggests a direct effect of gonadal steroids on reward processing, especially since reward learning was higher in the estrus stage of the cycle, when circulating hormone levels decline. This decreased focus on recent reward might also explain the reduction in trials completed, possibly reflecting attenuated motivation to obtain food reward (Supplementary Fig. 2). However, the observed effect on motivation may also be the result of cyclic changes in appetite (Tarttelin and Gorski, 1971).

The reduction in the value estimate of explore/exploit parameter β during estrus indicates that sexually receptive females chose more stochastically (i.e., shifting from exploitation to exploration of the response options) than during the non-receptive stages of the cycle, perhaps reflecting a general increase in exploratory behavior. At the same time, this increase in exploration may have resulted in reduced task engagement, leading to a decrease in the number of trials completed (Supplementary Fig. 2). Whether such cyclic changes in exploration have evolutionary advantage, for example by promoting search for a sexual partner, remains to be investigated.

Researchers are increasingly encouraged to include female animals in preclinical experiments, with the aim to increase the translational value of animal research. In this regard, our data provide further insight into the complexity of value-based decision making and its sex-specific modulation. Importantly, behavioral data from intact female animals should be properly controlled for the estrous cycle, since many behavioral tasks in neuroscience involve (food) reward, and are therefore subject to changes in value-based learning, motivation and appetite.

In sum, we provide direct evidence that reward learning, exploration and motivation, but not punishment learning and perseveration, fluctuate during the estrous cycle in female rats. Although cyclic changes in value-based decision making have been observed before, which computational components underlie these changes had not yet been elucidated. It is well known that gonadal steroids have widespread effects on the brain, including on the mesocorticolimbic dopamine system (McEwen and Alves, 1999), which is an important hub for value-based learning (Verharen et al., 2018). It is therefore likely that estradiol and progesterone affect reinforcement learning through corticolimbic mechanisms, to promote adaptive survival-directed behavior in females.

Author contributions

J.P.H.V. performed the behavioral experiments and analyzed the data. J.K. obtained the vaginal smears and determined the estrous cycle stage. L.J.M.J. and R.A.H.A supervised the experiments. All authors wrote the manuscript and have approved the final version of this paper.

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability

All data is publicly available at github.com/jeroenphv/EstrousCycle.

Acknowledgements

Funding was provided by the European Union Seventh Framework Programme (grant agreement number 607310; Nudge-It), and by the Consortium on Individual Development (CID), which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.psyneuen.2018.09.016.

References

Bari, A., Theobald, D.E., Caprioli, D., Mar, A.C., Aidoo-Micah, A., Dalley, J.W., 2010. Serotonin modulates sensitivity to reward and negative feedback in a probabilistic reversal learning task in rats. Neuropsychopharmacology 35, 1290–1301.
Cora, M.C., Kooistra, L., Travlos, G., 2015. Vaginal cytology of the laboratory rat and mouse: review and criteria for the staging of the estrous cycle using stained vaginal smears. Toxicol. Pathol. 43, 776–793.
Daw, N., 2009. Decision Making, Affect, and Learning: Attention and Performance XXIII Ch. 6.
Gershman, S.J., 2016. Empirical priors for reinforcement learning models. J. Math. Psychol. 71, 1–6.
Goldman, J.M., Murr, A.S., Cooper, R.L., 2007. The rodent estrous cycle: characterization of vaginal cytology and its utility in toxicological studies. Birth Defects Res. B Dev. Reprod. Toxicol. 80, 84–97.
Hendrick, V., Altshuler, L.L., Burt, V.K., 1996. Course of psychiatric disorders across the menstrual cycle. Harv. Rev. Psychiatry 4, 200–207.
Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., Daw, N.D., 2011. Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci. 14, 1250–1252.
Maia, T.V., Frank, M.J., 2011. From reinforcement learning models to psychiatric and neurological disorders. Nat. Neurosci. 14, 154–162.
McEwen, B.S., Alves, S.E., 1999. Estrogen actions in the central nervous system. Endocr.

Rev. 20, 279–307.
Posch, M., 1999. Win–stay, lose–shift strategies for repeated games—memory length, aspiration levels and noise. J. Theor. Biol. 198, 183–195.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II: Curr. Res. Theory 2, 64–99.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press.
Tarttelin, M.F., Gorski, R.A., 1971. Variations in food and water intake in the normal and acyclic female rat. Physiol. Behav. 7, 847–852.
Verharen, J.P.H., de Jong, J.W., Roelofs, T.J., Huffels, C.F.M., van Zessen, R., Luijendijk, M.C., Hamelink, R., Willuhn, I., den Ouden, H.E., van der Plasse, G., Adan, R.A.H., Vanderschuren, L.J.M.J., 2018. A neuronal mechanism underlying decision-making deficits during hyperdopaminergic states. Nat. Commun. 9.