Effects of dopamine D2 and opioid receptor antagonism on the trade-off between model-based and model-free behavior in healthy volunteers

Nace Mikus* ^{1,2}, Sebastian Korb ^{1,3}, Claudia Massaccesi ⁴, Christian Gausterer ⁵, Irene Graf ⁶, Matthäus Willeit ⁶, Christoph Eisenegger[†] ¹, Claus Lamm ¹, Giorgia Silani* ⁴, Chris Mathys* ^{2,7,8}

¹ Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Austria, ² Interacting Minds Centre, Aarhus University, Aarhus, Denmark, ³ Department of Psychology, University of Essex, Colchester, UK, ⁴ Department of Clinical and Health Psychology, Faculty of Psychology, University of Vienna, Austria, ⁵ FDZ-Forensisches DNA Zentrallabor GmbH, Medical University of Vienna, Austria, ⁶ Department of Psychiatry and Psychotherapy, Medical University of Vienna, Austria, ⁷ Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland, ⁸ Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy

* Senior authors contributed equally

[†] Deceased on the 27th of February 2017

• Correspondence to:

Nace Mikus, nace.mikus@univie.ac.at, Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Liebiggasse 5, 1010 Vienna,

Abstract

Our daily behaviour requires a flexible arbitration between actions we do out of habit and actions that are directed towards a specific goal. Drugs that target opioid and dopamine receptors are notorious for inducing maladaptive habitual drug consumption, yet how the opioidergic and dopaminergic neurotransmitter systems contribute to the arbitration between habitual and goal-directed behaviour is poorly understood. By combining pharmacological challenges with a well-established decision-making task and a novel computational model, we show that the administration of the D2 dopamine receptor antagonist amisulpride led to an increase in goal-directed or ‘model-based’ relative to habitual or ‘model-free’ behaviour, whereas the non-selective opioid receptor antagonist naltrexone had no appreciable effect. These findings highlight the distinct functional contributions of dopamine and opioid receptors to goal-directed and habitual behaviour and support the notion that D2 receptor antagonists promote stabilisation of goal-relevant information.

Introduction

Several theories of decision making postulate the existence of two distinct systems that drive our behaviour: a habitual system, which is automatic, reflexive and fast; and a goal-directed system, which is deliberative, reflective and effortful^{1–3}. This dichotomy of systems has a computational analogue in ‘model-free’ and ‘model-based’ decision-making models^{2,4}. A model-free agent simply selects actions that have led to rewarding outcomes in the past. This strategy is fast, computationally cheap, but can be inaccurate. A model-based agent uses an internal model of the environment to flexibly plan behavioural responses. This leads to more goal-oriented behaviour but is slower and relies on effortful cognitive control.

Studies investigating how individuals allocate control between the postulated two systems have shown that model-based control is increased when there is more to gain^{5,6}, and decreased when cognitive resources are scarce^{7,8}. Everyday behaviour can therefore be thought of as constant weighing of costs and benefits of applying model-based over model-free decision-making strategies^{6,9,10}. Failure to exert cognitive control over habitual urges in order to avoid negative outcomes has been suggested to be a hallmark of substance addiction^{2,11}. In support of this, studies show that decreased model-based control is linked to stimulant addiction¹² and seems to constitute a transdiagnostic dimensional trait related to compulsive behaviour across clinical^{12,13} and non-clinical populations¹⁴. In light of increasing deaths from drug overdoses^{15,16}, it is important to understand how different neurotransmitter systems affect the competition for cognitive resources when deciding between habitual and goal-directed actions.

Opiates, psychostimulants, and most other drugs of abuse increase the release of dopamine along the mesolimbic pathway^{17,18}, a circuit that plays a central role in reinforcement learning¹⁹. On top of this, the reinforcing properties of addictive drugs also depend on their ability to activate the μ opioid receptors^{20–22}. This invites the question whether the drift from goal-directed towards habitual control is enabled by dopamine and opioid receptors via a common neural pathway.

Yet, despite this question’s clinical relevance, causal investigations of the involvement of dopamine and opioid receptors in the trade-off between the two systems in humans are lacking. The evidence for the involvement of both dopamine and opioid receptors in forming and overcoming habits is restricted to preclinical studies on animals^{18,22} and to studies on patients with substance disorder^{23,24} with poor generalisability due to altered reward processing following chronic drug use²⁵. Moreover, pharmacological studies in healthy volunteers are limited to unspecific dopamine receptor agonists

and produced conflicting results^{26,27}. The variability in these results may be due to dopamine's involvement both in habit-formation as well as cognitive enhancement²⁸, whereby its effect likely depends on the binding site²⁹. Both D1 and D2 dopamine receptor families (called D1 and D2 in what follows) are involved in habit formation³⁰ but appear to have opposing contributions to goal-stabilisation. D1 dopamine receptors in the prefrontal cortex enable maintenance of goal-relevant information and working memory, while D2 dopamine receptor activity disrupts prefrontal representations and supports cognitive flexibility^{31–35}.

To address the limitations of previous studies, we directly compared the effects of the highly selective D2 dopamine receptor antagonist amisulpride with the opioid receptor antagonist naltrexone on the model-based/model-free trade-off in healthy volunteers. We hypothesized that both opioid and D2 antagonists would reduce habitual strategies and that blocking D2 receptors could also support model-based behaviour by stabilizing goal-relevant information³⁶.

We tested 112 participants with a deterministic version of the two-step task^{4,37}, at baseline and after administering amisulpride (N = 38, 400 mg), naltrexone (N = 39, 50 mg), or placebo (N = 35) in a randomized, double-blind, between-subject design (Fig. 1). The two-step task is a well-established incentivized paradigm, where participants need to make choices at the first stage, that influence their outcomes in the second stage. Each trial of the task starts in one of two possible first-stage states, each featuring a pair of spaceships (Fig. 2a). Upon choosing a spaceship, participants were taken to one of two planets in the second stage, where they encountered an alien that gave them points, which were converted to money at the end of the experiment. In each pair of spaceships, one spaceship flew to the red planet and one to the green. This deterministic mapping from spaceships to planets, as well as which spaceships were paired together, stayed constant throughout the task. However, the points received on each planet changed independently with a Gaussian random walk. When playing the game, participants could therefore simply repeat the choice of spaceship that previously led to positive outcomes, regardless of which planet it was associated with (model-free behaviour, Fig. 2b, c). Or they could attempt to remember which spaceship flew to which planet and choose the spaceship based on its associated planet (model-based behaviour, Fig. 2b, c). The task tried to mimic real-life decision-making dilemmas, where we can use our knowledge about the world (e.g., a potentially dangerous virus is circulating) to override a habitual action (e.g., shaking hands when greeting someone) with a more appropriate one (e.g., greeting with an elbow bump or a bow).

As an initial straightforward approach to dissociate model-based and model-free behaviour, we first examined how previous reward points affected the probability to stay with the previous choice. We then turned to computational modelling and explicitly modelled the relative weight, ω , of model-based compared to model-free behaviour and disentangled it from other processes that might affect participants' choices, such as exploration and reward devaluation. Based on our hypotheses, we expected that both amisulpride and naltrexone would increase the weighting parameter ω , but that this difference would be more pronounced for amisulpride (Fig. 2d).

To assess any effects of the two pharmacological compounds on cognitive control, we also collected data on a working memory task. Finally, since a systemic administration of pharmacological compounds does not on its own allow for definitive conclusions about the specific neurobiological mechanisms involved in the decision-making process, we additionally explored how four genetic markers of baseline dopamine function moderate potential drug effects. The corresponding results

will serve to direct future hypothesis-driven research. The data and the analysis scripts are available open-access at <https://github.com/nacemikus/mbmf-da-op.git>.

Results

Effects of dopaminergic and opioidergic antagonism on staying with previous choices

One way to quantify the trade-off between the two systems is by looking at the interaction effect of previous points on staying behaviour in trials where knowledge about the task structure is irrelevant (same first stage state as in the previous trial) with trials where it is relevant (different first stage states as in the previous trial). On average previous points increased the likelihood to stay with a previous choice ($\beta_{logodds} = 0.402$, 95% CI [0.251, 0.553], $P(\beta_{logodds} < 0) < 10e-3$, Fig. 3), however this was significantly less the case in trials with different vs. the same first stage states ($\beta_{logodds} = -0.204$ (95% CI [-0.263, -0.146], $P(\beta_{logodds} < 0) < 10e-3$). This indicates that participants often failed to consider the mapping from spaceship to planets when making choices in trials where the first stage state differed from the previous trial.

Compared to placebo, amisulpride significantly increased the difference between the effects of previous points on staying behaviour in different vs. same first state trials, as indicated by a significant four-way interaction ($\beta_{logodds} = 0.142$, 95% CI [0.059, 0.234], $P(\beta_{logodds} < 0) < 10e-3$, Fig. 3). Interestingly, when looking at each type of trial separately, we found that amisulpride both decreased the effect of previous reward points on the probability to stay when the first-stage state was the same ($\beta_{logodds} = -0.097$, 95% CI = [-0.175 -0.018], $P(\beta_{logodds} > 0) = 0.008$), as well as increased it with a trend when the first-stage state was different ($\beta_{logodds} = 0.046$ (95% CI [-0.007, 0.098], $P(\beta_{logodds} < 0) = 0.048$). For naltrexone, the log odds estimate of the four-way interaction coefficient was centred around -0.011 (95% CI [-0.105, 0.087], $P(\beta_{logodds} < 0) = 0.584$), indicating no marked difference between trial types.

Estimation of drug effects with computational modelling

We designed a novel computational model that reflects the structure of the task and embedded the model parameter estimation within a hierarchical Bayesian inference framework³⁸. We defined a model (M1) where, similarly to the computational models previously used with this task^{6,37}, the trade-off between the goal-directed and habitual components is captured by the weighting parameter ω , which embodies the degree to which the choice of participants on each trial is influenced by model-based ($\omega = 1$), or model-free ($\omega = 0$) subjective values of each of the spaceships. In this model, the subjective values of both the model-free and the model-based model components are defined as the last observed outcome following the choice of that spaceship. The crucial difference between the two components is that, whereas the model-free agent learns only by experiencing direct outcomes of spaceship selection, the model-based component always considers the deterministic mapping from spaceships to planets, and thus learns the subjective values of planets. We also include an inverse temperature parameter η (lower η indicates more explorative behaviour) and a *discounting parameter* γ that marks the degree of devaluation of non-chosen (and non-encountered) spaceships for the model-free component in each trial. The model was validated through parameter recovery and posterior predictive checks³⁸ and then compared to two other models commonly used in the field (see Methods and Supplementary Fig. 1).

Dopaminergic antagonism increases model-based relative to model free control

We first observed that the more model-based choices the participants made, the more money they earned ($r = 0.65$, 95% CI [0.53, 0.76]). This amounts to a validity check of the task, which was

designed to make cognitive control pay off (literally)³⁷. Importantly, we found that under amisulpride the difference in ω between the two sessions is higher than in the placebo group ($\beta_{ami}^{\Delta\omega} = 0.787$ 0.010, 95% CI [0.131, 1.510], $P(\beta_{ami}^{\Delta\omega} < 0) = 0.010$, Fig. 3a), with effect size $d = 0.758$, (95% CrI [0.126, 1.455]). In contrast, there was no session difference in ω between naltrexone and placebo ($\beta_{nal}^{\omega} = 0.238$, 95% CrI [-0.443, 0.856], $P(\beta_{nal}^{\omega} < 0) = 0.250$), with a marginally significant difference with a moderate effect size between the effects of the two compounds ($\beta_{ami-nal}^{\Delta\omega} = 0.578$, 95% CrI [-0.108, 1.337], $P(\beta_{ami-nal}^{\Delta\omega} < 0) = 0.046$, $d = 0.557$, 95% CrI [-0.104, 1.289]).

Effects of drugs on other model parameters, mood and working memory

We found no evidence that either the blockade of dopamine D2 or opioid receptors influenced the devaluation parameter γ (Fig. 3b, d, $\beta_{ami}^{\gamma} = 0.01$, 95% CI [-1.00, 1.00], $\beta_{nal}^{\gamma} = 0.07$, 95% CI [-0.88, 1.02]). However, we found some evidence for amisulpride effects on the inverse temperature parameter η (Fig. 3c, d, $\beta_{ami}^{\eta} = -0.33$, 95% CI [-0.67, 0.02], $P(\beta_{ami}^{\eta} > 0) = 0.034$, $d = -0.38$, 95% CI [-0.78, 0.03]). This would imply that amisulpride increases “explorative” choices or choices that the model would not predict based on estimated action values. To verify if our model performs worse for the amisulpride group we looked at how the pharmacological treatment predicts out-of-sample prediction accuracy and found no differences between groups (Supplementary Fig. 1f, Supplementary Table 2). Note also that the differences between sessions in ω and η are not negatively correlated ($r = 0.12$, 95% CI [-0.06, 0.30]), suggesting that the two effects are not related to each other.

To examine whether working memory capacity was affected by our drug manipulation, we analysed performance in the Reading Span task³⁹ and found no evidence for drug effects on proportion of correct recalls ($\beta_{ami}^{wm} = -0.02$, 95% CI [-0.07, 0.02], $\beta_{nal}^{wm} = -0.03$, 95% CI [-0.08, 0.02]). The increase in model-based behaviour could also be due to general effects on mood. We assessed participants’ mood on the day of drug administration twice, at drug intake and 3 hours later (Supplementary Table 3). A Bayesian linear model, predicting mood scores for positive and negative affect (PANAS)⁴⁰, with varying intercepts for each participant, showed that in the placebo group there was a flattening of positive ($\beta_{time}^{+} = -0.38$, 95% CI [-0.58, -0.18]) and, to a lesser degree, negative ($\beta_{time}^{-} = -0.2$, 95% CI [-0.46, 0.06]) mood after 3 hours of testing (for the full table see Supplementary Table 4). The flattening in both affective domains was even more pronounced in the naltrexone group ($\beta_{time:nal}^{+} = -0.34$, 95% CI [-0.63, -0.04], $\beta_{time:nal}^{-} = -0.45$, 95% CI [-0.83, -0.07]), but was comparable to placebo in the amisulpride group ($\beta_{time:ami}^{+} = -0.2$, 95% CI [-0.48, 0.09], $\beta_{time:ami}^{-} = -0.12$, 95% CI [-0.5, 0.25]). Finally, we ran a model predicting differences in ω between sessions, with working memory, mood and their interactions with the drug administration as predictors and found no effects (Supplementary Table 5), suggesting that the effects of amisulpride on ω were not due to effects of the drug on mood or working memory.

Explorative analysis of the interaction effects of pharmacological administration with genetically determined baseline dopamine function

In line with previous studies using similar paradigms, we used the *COMT* single nucleotide polymorphism (SNP) as a marker for prefrontal dopamine function, and *DARPP-32* SNP as a marker of striatal D1 receptor function^{41–43}. We also investigated two genetic markers for striatal dopamine levels, namely the *Taq1a* SNP and the *DAT1* polymorphism, a 40 base-pair variable number tandem repeat polymorphism (VNTR) of the dopamine transporter^{44–46} (see Supplementary Note 2 for details on genotypes and how they relate to striatal and prefrontal dopamine function).

To estimate the moderating effects of the genotype variables, we ran again the hierarchical estimation of the parameters using shrinkage mixture priors (spike-and-slab) to regularize the parameter estimation⁴⁷. The spike-and-slab prior pulls small effects towards zero and weakly regularizes non-zero effects^{48,49}. To get a less conservative estimate of the effect, we ran a simple linear model predicting estimated changes of ω from one session to the next from drug and genotype interactions (Supplementary Fig. 2 a, b). We find that COMT moderated the effects of amisulpride ($d = 0.597$, 95% CrI [0.155, 1.049], $P(d < 0) = 0.005$, Supplementary Fig. 2c), meaning that participants with genetic correlates of higher prefrontal dopamine were more likely to increase their model-based control following amisulpride administration. For naltrexone, we found a moderating effect of the Dat1 polymorphism ($b = 1.077$, 95% CrI [0.108, 2.038], $P(d < 0) = 0.015$), and, to a lesser degree, of the other genotypes related to striatal dopamine function (Supplementary Fig. 2d). Note however that all the 95% posterior distributions of effect sizes from shrinkage priors contained zero (Supplementary Fig. 2b).

Discussion

Drugs that stimulate opioid or dopamine neurotransmitter systems can lead to compulsive drug-taking, which can result either from increased reliance on habits or from failures to exert cognitive control over habitual urges. Here, we asked whether pharmacologically blocking opioid and dopamine receptors promotes the use of goal-directed over habitual control in healthy volunteers. Using the two-step task and a novel computational model, we found that blocking D2-like dopamine receptors increases the weight on model-based/model-free control, whereas blocking opioid receptors has no appreciable effect, with a marginally significant difference with a moderate effect size between the effects. The results from the computational model mirrored those obtained when analyzing the effects of previous points on staying with prior choices. These findings highlight the differential roles of the opioid and dopamine receptors in the model-based/model-free trade-off.

Prior research on dopamine's involvement in the arbitration between the two systems has produced inconsistent results. For instance, elevated dopamine levels, either at baseline or following pharmacological treatments such as L-dopa, are related to increased model-based behaviour in some studies^{26,50}, but show no correlation or even opposite effects in others^{27,51}. This suggests that the effect of dopamine might depend on the relative contributions of the specific dopamine receptor subtypes or on the baseline differences in dopamine levels across different brain areas. It is well-established that dopamine enables cognitive control, mainly through prefrontal D1 receptors^{32,52}. Striatal dopamine on the other hand is believed to control the gating of working memory content and therefore facilitates goal-relevant updating of information in the prefrontal cortex^{35,53}. In the striatum, D2 receptor expressing cells are known to exert a tonic inhibitory effect on prefrontal and local circuits^{54,55}. Higher D2 receptor activity disinhibits the cumulative basal ganglia output and promotes behavioural flexibility³⁵. In line with this, D2 antagonists impair behaviour in tasks that require constant attentional set shifting⁵⁶, but might improve performance in tasks where prefrontal goal representations are required⁵⁷, such as the two step-task. For instance, D2 antagonism improves the ability of participants in certain cognitive tasks that require manipulation of task-relevant information in the working memory^{36,53}, but does not improve working memory capacity, or memory retrieval^{36,58,59}. This is supported by our data since we found no effect of amisulpride on working memory performance. Furthermore, an exploratory analysis of genetic data indicated that the effect of amisulpride might be moderated by prefrontal dopamine availability, which is known to be crucial in enabling model-based control^{41,60}. This suggests that the D2 receptor antagonists' effect on model-

based behaviour is likely due to inhibiting disruptions of task-relevant prefrontal representations rather than to an increased capacity for cognitive control per se.

More recently, it has also been suggested that the striatal mediation of cognitive control reflects a willingness to exert cognitive effort, through a comparison of potential gains of applying cognitive control against the costs of doing so^{29,61}. For example, a recent study showed that striatal dopamine synthesis capacity, measured with positron emission tomography, was related to higher motivation to exert cognitive effort⁶². Furthermore, in participants with low striatal presynaptic dopamine levels, sulpiride, an antipsychotic with a similar binding profile as amisulpride, increased the perceived benefits vs. costs of cognitive effort, an effect that was comparable to that of methylphenidate, a drug that increases dopamine availability in the synaptic cleft. D2 dopamine receptors are known to regulate dopamine reuptake, release, and synthesis through presynaptic auto-receptors⁶³. Blocking D2 receptors could therefore boost the motivation to apply model-based control through increasing striatal dopamine levels. Interestingly, a recent study found that L-DOPA, a drug that increases presynaptic dopamine, did not increase model-based behaviour, but led to a general reduction of the effect of previous reward on choice, which was explained as reduced model-free behaviour, and higher noise.

In line with this, another explanation of amisulpride's effects on the weighting parameter ω in our study is that they arise as compensation for disturbed model-free control. In previous studies, D2 antagonists led to poorer performance in simple reinforcement learning tasks^{64,65}, reduced cue responding following Pavlovian conditioning, as well as preference for immediate rewards in a delayed discounting task⁶⁶. In a task completed by the same cohort of participants prior to the two-step task, amisulpride reduced the willingness to exert physical effort to obtain immediate primary rewards⁶⁷. In the present study, we found that amisulpride decreased the effect of previous points on staying behaviour in trials where the first-stage state was the same as in the previous trial, and therefore remembering the task structure was unnecessary. However, amisulpride increased the effect of previous points on choice in trials with different first stage states, suggesting that it did not disrupt learning overall. Another explanation for this is that amisulpride increased the capacity (or willingness) for model-based behaviour, but also increased choice exploration. Previous studies have implicated D2 receptors in exploration^{46,68} and we found some evidence of a slightly decreased inverse temperature parameter (i.e., more 'noisy' or explorative behaviour) in participants who received amisulpride. Furthermore, the drug effects on the exploration parameter and on the model-based/model-free weight were not correlated. Therefore, the increased exploration does not explain increased model-based/model-free weight and might reflect an orthogonal neurobiological effect. Finally, participants could show reduced model-free learning that would reduce the effect of previous points on choice in trials with the same first-stage states, and at the same time reduce the misleading model-free action-outcome coupling that disrupts the model-based choice selection in trials where consecutive first states are different.

The lack of a pronounced effect of naltrexone on model-based/model-free behaviour is unlikely to be due to issues of dosage or timing, since 50 mg of naltrexone lead to a ~90% of μ receptor occupancy even 48 hours post intake⁶⁹, with μ receptors being the primary (but not only) opioid receptor type that the drug binds to. One possibility is that opioid receptors are not crucial for model-free learning required in this task. This is in opposition to previous studies showing that acute administration of naltrexone, comparably to amisulpride, causes a reduction of cue responsivity and reward impulsivity⁶⁶, decreases effort to obtain immediate primary rewards⁶⁷, and decreases the wanting of

rewards⁷⁰. Similarly, preclinical studies in rats show that naltrexone and naloxone, another opioid antagonist, decreased sucrose reinforced place preference^{71,72}. Furthermore, both naltrexone and naloxone are used to reduce craving in patients with addiction²⁴. However, there is also evidence that opioid receptors are important for goal-directed behaviour. In particular, a study in rats showed that opioid receptor blockade leads to decreased sensitivity to reward value and accelerated habitual control of actions⁷³, and in a recent neurochemical imaging study in humans μ opioid receptor availability correlated with goal-directed behaviour in a loss-only version of the two-step task⁵¹. In fact, opioid agonists can in some cases lead to increased performance in cognitive control tasks⁷⁴. One explanation of our results is that naltrexone simply reduced the intrinsic value of reward and therefore decreased the motivation to exert cognitive effort. This would be in line with the observed flattening of both positive and negative affect after naltrexone administration compared to placebo. Finally, the lack of effects of naltrexone on the model-based/model-free trade-off might simply be too weak to detect in this study. Exploratory analysis of genetic data tentatively suggests that the effects might only be present in individuals with higher striatal dopamine levels, supporting the notion that habit formation through opioid receptors depends on striatal dopaminergic circuits.

An important limitation of the experimental approach used here is that it rests on the assumption that there are only two ways of learning, which moreover can be distinguished clearly. This assumption has been questioned in the reinforcement learning literature^{75,76}, and mirrors the scepticism of the two-system division of decision making in cognitive psychology⁷⁷. Although we made sure that participants understood what the task rules were and how they could maximize their gains, we cannot exclude the possibility that participants searched for alternative models to obtain rewards, such as different pressing patterns or simply favouring one stimulus over the other⁷⁵. In fact, the slightly increased exploration parameter in the amisulpride group could be due to participants' increased exploration of this model space.

In conclusion, we provide a first comparison of the contributions of the dopamine D2 and opioid receptors to arbitration between the model-based/model-free systems. Our results highlight the role of D2 dopamine antagonists in promoting goal-stability when alternative habitual choices are available. Furthermore, we show that opioid blockers in general do not promote model-based behaviour. These findings are a step forward in understanding how neuromodulators control the arbitration between habitual and goal-directed decision-making systems, which can on the long run be crucial for developing targeted pharmacological treatments for addiction and other disorders of compulsivity.

Methods

Procedure

The study took place in the Department of Psychiatry and Psychotherapy at the Medical University of Vienna. After initial online screening, the participants were invited for a first visit, where they underwent a physical examination (electrocardiogram, hemogram) and a psychiatric screening before being subjected to the two-step task that consisted of a 25 - minute automated training period followed by 200 trials of the task (see Figure 1 for the study outline). Participants eligible for the study were invited back to the clinic approximately 3 weeks after their first visit.

In their second visit participants received either placebo, 400 mg of amisulpride, or 50 mg of naltrexone. After a waiting period of 3 h, participants solved two additional tasks (explained elsewhere, see Korb et al., 2020) before performing the two-step task (on average 4 h 36 min, SD =

22 min, after the pill intake) as well as a working memory task. Waiting times and the doses of drugs were chosen based on previous pharmacological studies with the same compounds⁶⁶. To ensure comparable absorption, participants were asked to come with an empty stomach and received a standardized meal before pill intake. Participants filled out a questionnaire assessing mood and side effects right after pill intake and 3 hours later. There were no profound differences in side effects across the three drug groups, apart from a trend level effect of amisulpride on tiredness (Supplementary Fig. 3). Blood plasma levels confirmed amisulpride levels above 212.6 $\mu\text{g/l}$ (mean (sd) = 548.8 (96.9)), in all participants of the amisulpride group.

Participants

Data was collected from 120 volunteers and constituted a subset of volunteers involved in a larger study (Korb et al., 2020). All participants were assigned an initial screening code and a subject ID number. For six participants the task data from the first session were lost due to failures in assigning screening codes from the first session to subject ID numbers in the second. For two participants the task data from the second session were lost owing to technical issues, resulting in 112 participants included in the main analysis (see Supplementary Table 10 for exact subject numbers for each analysis). All participants had no history of drug abuse or other psychiatric disorders and were matched in age, sex, and BMI (Supplementary Table 8). The study was approved by the Ethical Committee of the Medical University of Vienna (EK N. 1393/2017) and in line with the Declaration of Helsinki⁷⁸. Participants received a monetary compensation of 90€ plus the extra money earned in the task.

Genotypic analysis

Peripheral blood was collected by lancet and stored on Whatman FTA micro cards (Sigma-Aldrich). DNA was extracted using the QIAamp DNA Mini kit (Qiagen, Hilden, Germany). The VNTR polymorphism in the DAT1 gene was investigated by PCR with 5'-fluorescent-dye-labeled forward primer and automated detection of PCR products by capillary electrophoresis (details of the procedure provided in the supplement). The single base primer extension (SBE) method also known as SNaPshot minisequencing was applied for the typing of single nucleotide polymorphism (SNP) variants (details provided in supplement). Accordingly, five informative SNPs [ANKK1 (rs1800497), BDNF (rs6265), CDH13 (rs3784943), OPRM1 (rs1799971) and PPP1R1B (rs907094)] were analysed simultaneously applying a multiplex strategy for PCR and SNaPshot minisequencing of purified PCR products. Typing of Val158Met variants (rs4680) in the COMT gene was carried out separately, applying a singleplex approach for PCR and SNaPshot. The OPRM1 and BDNF SNPs were not included in the analysis in this task. Genetic data from three participants were lost. This led to the group distributions depicted in Supplementary Table 9. For details on the analysis see Supplementary Note 3).

Task design

In the task, participants made an initial choice at stage one which took them to one of the two possible states ('planets') in stage two. There were two possible states in the first stage, each featuring a pair of spaceships. Each of the spaceships flew deterministically to one of the two planets where they encountered an alien who gave them either positive or negative points. The points in each planet changed according to a Gaussian random walk in the interval $[-4, 5]$, rounded to whole numbers. Importantly, since both first-stage states led to the same two possible second-stage states, participants could transfer knowledge from the pair of spaceships in one first-stage state to the pair in the other. According to the conventional definition adopted here, a completely model-free agent

relies entirely on its direct experience and will choose a spaceship based only on its reward history in previous trials featuring the same first-stage state, regardless of their experience with the other pair of spaceships. A completely model-based agent however will use the causal structure of the task to update the value of the spaceships based on which planet they fly to. This means that knowing that spaceship A and C fly to the same planet (but appear in different first step states) enables the model-based agent to learn about spaceship C by experiencing the outcome of its choice of spaceship A.

Participants had 2000 ms to choose in the first step and then again 2000 ms to press the space bar once they had encountered the alien. Each acquired point was translated to 4 Eurocents and added to the overall compensation of the participant at the end of the experiment. The story and a thorough training session were employed to increase the comprehension of the task, as done before⁷⁹. To avoid all participants behaving in a model-based way⁷⁵ we increased the difficulty of the task by dynamically changing the drift of the reward-determining Gaussian walks. The Gaussian random walks therefore had various drift rates (0.5, 1, and 2). We generated one trajectory and shuffled it around to create two different testing sessions of similar difficulty. Each participant first went through an automated rigorous explanation of the task followed by 25 practice trials before completing 200 trials of the task.

Behavioural analysis

To regularize our inference, we used a hierarchical Bayesian approach in all our analyses. We report estimates in parameter estimation space and indicate the precision of our estimates with credible intervals^{38,80}. We also report the proportion of the credible interval that is above or below zero. Model-agnostic analysis of behaviour focused on predicting the probability of staying with the previous choice (choose the spaceship that flies to the same planet as in the previous trial). Behaviour was analysed using the *brms* package in R⁸¹ which employs the probabilistic programming language Stan⁸². We fit a binomial model that predicted staying with the previous choice from reward obtained in the previous trial, modulated by the previous state (same or different), drug treatment, and session. The effects of session, previous state, and previous points were drawn from a multivariate normal distribution (were considered as correlated random effects). We report 95% credible intervals of estimates on the log-odds scale as well as in terms of odds (the ratio between probabilities of staying and switching) for a more comprehensive summary of the results. Parameters were estimated with 4 chains, each 3000 iterations (1000 warmup), with priors listed in Table 1.

Table 1: Prior distributions for the behavioural analysis

Standard Deviations	$\sigma \sim \text{HalfCauchy}(0, 2)$
Regression Coefficients	$\beta \sim N(0, 3)$
Intercept	$\beta_0 \sim \text{Student}(3, 0, 10)$
Prior for the correlation matrix	$R \sim \text{LKJcorr}(2)$

The number of participants used in the analysis depend on whether both sessions were included and whether genetic data were used (see Supplementary Table 10 for an overview).

Computational Models

We defined three computational models that define the subjective values Q of participants' action a , in trial t , with a first stage state s , as the weighted average of model-based (Q_{MB}) and model-free (Q_{MF}) subjective values.

$$Q(a, s, t) = \omega Q_{MB}(a, s, t) + (1 - \omega) Q_{MF}(a, s, t)$$

where ω is the weighting parameter (larger influence on choice of model-based values is indicated by ω being close to 1 and model-free control by ω being close to 0). The subjective values are then mapped on to probability of choosing a , and not a' , with the soft-max transformation:

$$P(a, s, t) = \frac{1}{1 + e^{-\eta*(Q(a,s,t)-Q(a',s,t))}}$$

where η is the *inverse temperature* parameter, that determines the stochasticity of choices and the exploration-exploitation trade-off.

Deterministic learning model (M1)

Because the outcomes at the second stage are determined by Gaussian random walks, the objectively best prediction by the agent is the last encountered outcome. We defined a model (M1), where the model-free subjective value of each of the spaceships is the last encountered outcome following the choice of that spaceship, and the model-based subjective value of each spaceship is the subjective value of the last encountered outcome following the planet that this spaceship flies to. This means that the update equation of the model-free agent after receiving outcome $r(s_2, t)$ following an action a , in trial t , with first state s_1 , is defined as

$$Q_{MF}(a, s_1, t) = r(s_2, t).$$

In the same trial, the value of the for the unchosen actions across both first level states are shrunk towards 0 with forgetting parameter $\gamma \in [0,1]$:

$$Q_{MF}(a', s_1, t) = (1 - \gamma) Q_{MF}(a', s_1, t - 1),$$

$$Q_{MF}(a, s_1', t) = (1 - \gamma) Q_{MF}(a, s_1', t - 1),$$

$$Q_{MF}(a', s_1', t) = (1 - \gamma) Q_{MF}(a', s_1', t - 1),$$

where a' is the unchosen action and s_1' is the unencountered state. The model-based agent, on the other hand, after receiving outcome $r(t)$ following an action a , in trial t , with first state s_1 , updates not only the experienced first stage state, but also the unencountered first stage which would have led to the same outcome:

$$Q_{MB}(a, s_1, t) = Q_2(s_2, t) = r(s_2, t),$$

$$Q_{MB}(a, s_1', t) = Q_2(s_2, t) = r(s_2, t),$$

where, $Q_2(s_2, t)$ is the subjective value of the second stage state that action a deterministically leads to.

Dual-system reinforcement learning models (M2 and M3)

We compared our model M1 to two versions of a dual-system reinforcement learning model inspired by Kool et al (2016, 2017). In these models, the model-free agent learns the subjective values of spaceships and planets through a temporal difference-learning algorithm^{9,83}. The model-free agent was defined by 3 free parameters: the learning rate at the first stage (α_1), and the second stage (α_2), where the eligibility trace (λ) determines the degree to which the outcome at the second stage retrospectively transfers to the first stage. In simple terms, the model increases (or decreases) the subjective value of an action at stage 1 proportionally to how positively (or negatively) surprising the outcome was, but discounted by the learning rate that describes the contributions of previous outcomes of that specific action. Conversely, the model-based agent is aware of the structure of the task. Specifically, in trial t , with first stage state s_1 we define the model-free subjective value of action a as

$$Q_{MF}(a, s_1, t) = Q_{MF}(a, s_1, t) + \alpha_1 \delta_1$$

$$\delta_1 = Q_2(s_2, t) - Q_{MF}(s_1, a, t)$$

and the model-based subjective values of action a as

$$Q_{MB}(s_1, a, t) = T(s_1, a) * Q_2(s_2, t)$$

where $Q_2(s_2, t)$ is the subjective value of the second stage s_2 at time t and $T(s_1, a)$ is the transition matrix from first to second stage states. The agents learn the values of the planets, Q_2 , by

$$Q_2(s_2) = Q_2(s_2) + \alpha_2 \delta_2$$

$$\delta_2 = r - Q_2(s_2)$$

$$Q_{MF}(s_1, a) = Q_{MF}(s_1, a) + \lambda \alpha_1 \delta_2$$

The learning rates at the two stages were either different (M2), or the same (M3).

Model Estimation

We estimated the model parameters for both sessions in one hierarchical (multilevel) Bayesian model. This approach pools information across different levels (drug groups, and participants) and thus leads to more stable individual parameter estimates⁸⁴, reduces overfitting⁸⁵, and enables us to estimate in one model both individual and group level parameters as well as differences between sessions⁸⁶. Models were implemented in Stan⁸² using R as the interface. Stan uses a Markov Chain Monte Carlo sampling method to describe posterior distributions of model parameters. We ran each candidate model with four independent chains and 3000 iterations (1000 warm-up). Convergence of sampling chains was estimated through the Gelman-Rubin $\hat{R}$ statistic⁸⁷, whereby we considered $\hat{R}$ values smaller than 1.01 as acceptable.

For all subject level parameters (e.g., ω) we drew both the baseline (ω_0) and the session difference ($\Delta\omega$), from a multivariate Gaussian prior. Specifically, for model M1:

$$\begin{pmatrix} \omega'_0 \\ \gamma'_0 \\ \eta'_0 \\ \Delta\omega' \\ \Delta\gamma' \\ \Delta\eta' \end{pmatrix} \sim MVNormal \left(\begin{pmatrix} \mu_{\omega'_0} \\ \mu_{\gamma'_0} \\ \mu_{\eta'_0} \\ \mu_{\Delta\omega'} \\ \mu_{\Delta\gamma'} \\ \mu_{\Delta\eta'} \end{pmatrix}, S \right)$$

where S is the covariance matrix, that was factored into a diagonal matrix with standard deviations and the correlation matrix $R^{81,85}$. The prime denotes the parameters in estimation space. The parameters that were fully constrained (i.e., ω , α_1 , α_2 , λ , γ) were estimated in the inverse probit space and the parameters that only had a lower bound (i.e., η) were estimated in log space. The hyper-priors for all group-level means were weakly informative, $\mu_{\omega'} \sim N(0,1)$, the prior for group-level standard deviations were $\sigma_{\omega'} \sim \text{HalfNormal}(0,1)$, and the prior for the correlation matrix was $R \sim \text{LKJcorr}(2)$. The parameters in the second session were therefore defined by their estimated baseline and difference between the sessions. For instance, the model-based weight ω in the second session for participant i was defined as:

$$\omega(i) = \phi(\omega'_0(i) + \Delta\omega'(i))$$

where ϕ is the cumulative distribution function of the standard normal distribution (inverse of probit). Effect sizes (d) are calculated by normalizing the relevant regression coefficients by the pooled standard deviation (square root of the sum of all relevant variance components^{88,89}). In models where there are no random effects, this reduces to a Cohen's d . In particular, we defined the effect sizes by dividing the estimated difference between group means by the square root of the sum of the variance of both the baseline ($\sigma_{\omega'_0}^2$) and the session difference ($\sigma_{\Delta\omega'}^2$). We used the same procedure for all three models.

Model Comparison and Validation

We used the trial-based Leave-One-Out Information Criterion (LOOIC) to compare the three models using the `loo` package in R⁹⁰. The LOOIC estimates out-of-sample predictive accuracy of each trial and is more informative than simpler point-estimate information criteria used commonly (such as the Akaike information criterion). Lower LOOIC scores indicate better prediction accuracy out of sample. Additionally, we compared models with bootstrapped pseudo Bayesian Model Average relative weights of the models that reflect the posterior probability of each model given the data⁹¹. To validate the novel model M1, we first used the posterior means of the estimated parameters to simulate behaviour in both sessions. To see how well we can retrieve the model parameters we reran the parameter estimation on the simulated behaviour with the same model (Supplementary Fig. 1d). We also ran the same analysis of staying behaviour on this synthetic dataset and reproduced the behavioural plots from Fig. 3 (compare to Supplementary Fig. 1e). We used the same logistic hierarchical Bayesian model to statistically evaluate the crucial aspects of the behavioural analysis (Supplementary Fig. 1e, compare to Fig. 3c). To get the posterior predictive accuracy of the model we predicted the choice on each trial for each participant for 8000 samples drawn from the posterior distribution and then calculated the average accuracy for each participant.

Estimation of the effects of the pharmacological treatment and genotypes

To statistically evaluate the effect of our treatment on all the parameters in the model, we included two regression terms when defining the group-level means of the difference between sessions:

$$\begin{pmatrix} \omega'_0 \\ \gamma'_0 \\ \eta'_0 \\ \Delta\omega' \\ \Delta\gamma' \\ \Delta\eta' \end{pmatrix} \sim \text{MVNormal} \left(\begin{pmatrix} \mu_{\omega'_0} \\ \mu_{\gamma'_0} \\ \mu_{\eta'_0} \\ \mu_{\Delta\omega'} + \beta_{Ami}^{\omega} * X_{Ami} + \beta_{Nal}^{\omega} * X_{Nal} \\ \mu_{\Delta\gamma'} + \beta_{Ami}^{\gamma} * X_{Ami} + \beta_{Nal}^{\gamma} * X_{Nal} \\ \mu_{\Delta\eta'} + \beta_{Ami}^{\eta} * X_{Ami} + \beta_{Nal}^{\eta} * X_{Nal} \end{pmatrix}, S \right)$$

where β_{Ami}^* and β_{Nal}^* are coefficients drawn from prior distribution $N(0,1.5)$ and X_{Ami} and X_{Nal} are dummy variables for the two drugs. To see which of the parameters are affected by the drug treatment we included group-level effects on all parameters.

To estimate the effects of the genotype on model-based behaviour, we again re-estimated the model adding the group level effects for each genotype to the session 1 estimates and interaction effects to the difference between session. All genotype variables were coded as binary and centred, except for the *COMT* which was coded as continuous and standardized. The hyper-mean of the model-based weights for the baseline was defined as:

$$\mu_{\omega_0}^{gen} = \mu_{\omega_0'} + \beta_{COMT} * X_{COMT} + \beta_{Taqla} * X_{Taqla} + \beta_{Dat1} * X_{Dat1} + \beta_{DARPP} * X_{DARPP}$$

And for the session difference it was defined as:

$$\begin{aligned} \mu_{\Delta\omega'}^{gen} = & \mu_{\Delta\omega'} + (\beta_{COMT*Ami} * X_{COMT} + \beta_{Taqla*Ami} * X_{Taqla} + \beta_{Dat1*Ami} * X_{Dat1} + \beta_{DARPP*Ami} \\ & * X_{DARPP} + \beta_{Ami}) * X_{Ami} \\ & + (\beta_{COMT*Nal} * X_{COMT} + \beta_{Taqla*Nal} * X_{Taqla} + \beta_{Dat1*Nal} * X_{Dat1} + \beta_{DARPP*Nal} \\ & * X_{DARPP} + \beta_{Nal}) * X_{Nal} \end{aligned}$$

where X_{gene} are genotype variables and β_{gene} and $\beta_{gene*drug}$ are coefficients drawn from a spike-and-slab prior, a prior that is a discrete mixture of a peaked prior around zero and a vague proper normal prior^{48,49}. Regression coefficients with weak effects are pulled towards zero, and the coefficients that are above zero are estimated with adaptive shrinkage. Specifically, we define the priors as

$$\beta_j \sim \gamma_j * N(0, \tau_j^2) + (1 - \gamma_j) * N(0, 0.001),$$

$$\gamma_j \sim U(0,1),$$

$$\tau_j^2 \sim \text{Inverse Gamma}(0.5, 0.5)$$

We only looked at the effect of genotypes on model-based weights. The effect sizes are calculated again by dividing the posterior distributions of regressors with the squared root of the sum of both the variance of the baseline ω_0 as well as the session difference. To get a less conservative estimate of the effects, we ran another model predicting estimated differences in ω across sessions from the drug variables and their interactions with the four genotype variables. Here, we standardized the response variable and used weakly informative priors ($N(0,1)$) on the regression coefficients and fat-tailed priors on the variance (*HalfCauchy*(0,2)). Effect sizes are calculated by dividing the posterior distributions of coefficients by the estimated standard deviation.

Reading Span Task

We used an automated version of the Reading Span Task, where in each block participants saw 2 to 6 words serially presented that they had to recall by the end of the trial in any order. The words were interlaced with sentences that participants were instructed to judge as either making sense or not³⁹. Participants played 15 blocks. Correctly recalled items were calculated as a proportion within the block and then averaged across blocks. The effect of the drug treatment was calculated with a Bayesian linear model where the mean score was predicted by drug treatment.

Acknowledgements

The study was supported by the Vienna Science and Technology Fund (WWTF) with a grant (CS15-003) awarded to Giorgia Silani and Christoph Eisenegger and a grant (VRG13-007) awarded to Christoph Eisenegger and Claus Lamm. We thank Prof. Boris Quednow for his advice on the study design. This work would not be possible were it not for the students who have carried out the data collection: Mani Erfanian Abdoust, Anne Franziska Braun, Raimund Bühler, Lena Drost, Manuel Czornik, Lisa Hollerith, Berit Hansen, Luise Huybrechts, Merit Pruin, Vera Ritter, Frederic Schwetz, Conrad Seewald, Carolin Waleew, Luca Wiltgen, Stephan Zillmer. We thank Catherine Hartley for sending us her version of the task implemented in matlab, and for allowing us to use the stimuli. We are grateful to Wouter Kool for making his version of the task implemented in JsPsych freely available on github. We also thank Lei Zhang and Lukas Lengersdorff for helpful discussions on the statistical procedures used in the manuscript.

Figures:

Fig. 1 | Study Procedure. After an initial online screening, participants were invited to the lab for a first visit (Session 1), where they were subjected to a medical screening, before playing the two-step task for the first time. If they fulfilled the study criteria, they were invited for another visit (Session 2), where they received either 400 mg of amisulpride, 50 mg of naltrexone, or placebo (mannitol). After 180 minutes of waiting time, participants started with the test battery. Approximately 270 minutes after drug intake, participants performed the two-step task the second time, followed by a Reading Span (working memory) task, and a blood draw to determine amisulpride serum levels.

Fig. 2 | Task Design and Hypotheses. **a**, In the first stage of the trial, the participants were presented with one of the two pairs of spaceships. Each of the spaceships flew to one of two planets in the second stage, where they encountered an alien that gave them points. The transition from each of the four spaceships to the planets was deterministic and stayed the same throughout the experiment. The points each alien gave changed independently according to a discretized Gaussian random walk with bouncing boundaries at -4 and 5. **b**, Behaviour in the task showcasing trials where the previous first-stage state (spaceship pair) was the same (trial 2) and trials where it was different (trial 3). High number of points should encourage choosing the spaceship that flies to the same planet, what we term as “staying with the previous choice”. Note that in the featured example, the participant after receiving -1 point in the trial 1 opted against staying with the previous choice in the next trial, and after receiving 2 points in trial 2 opted for staying with the previous choice, by choosing the spaceship, that flew to the same planet. **c**, Model-free behaviour is defined by simple choosing the spaceship if it previously led to rewarding outcomes. In contrast, when behaving in a model-based manner, we choose the spaceship based on the planet it flies to, keeping the mapping from spaceships to planets in mind. **d**, We hypothesized that both amisulpride and naltrexone will increase the weight on model-based vs. model-free behaviour, however amisulpride would on top of that also support the model-based strategy.

Fig. 3 | Behavioural analysis. **a**, We used a hierarchical Bayesian logistic regression model to analyse how the probability to stay with the previous choice depended on previous points, previous first-stage state, session, and drug administration and their interaction. We allowed the intercept and the slopes for previous first-stage state and session to vary by participant (full table of coefficients in Supplementary Table 1). Points and error bars depict mean proportions and standard errors in trials where the participants stayed with their previous choice for each previous point averaged within session and drug group. Lines and ribbons depict the mean estimates and 80% credible intervals (CI). When encountering a trial with the same first-stage state as in its preceding trial, participants were 1.495 times (95% CI [1.286, 1.739], $P(\beta_{\log odds} < 0) < 10e-3$) more likely to stay with their previous choice for each additional previous point ($\beta_{\log odds} = 0.402$, 95% CI [0.251, 0.553], $P(\beta_{\log odds} < 0) < 10e-3$). In contrast, in trials where the first-stage state was different from the preceding trial, the odds (on the logarithmic scale) of repeating the previous choice of spaceship were reduced, with each additional point earned in the previous trial, by -0.204 (95% CI [-0.263, -0.146], $P(\beta_{\log odds} > 0) < 10e-3$), and participants were only 1.219 times (95% CI [1.053, 1.410], $P(\beta_{\log odds} < 0) < 10e-3$) more likely to stay with their choice for each additional previous point. This indicates that participants often failed to consider the mapping from spaceship to planets when making choices in trials where the first-stage state differed from the previous trial. The effects of the drugs can be seen by the different slopes of previous points in the two sessions depicted for both trial types. **b**, Differences in staying behaviour between sessions, binned into 5 different reward levels for clarity. Means with standard errors overlayed with means and 80% CI of estimated posterior distributions. **c**, Means with 80% and 95% CIs of effect sizes (in logodds space) of selected regression coefficients of the hierarchical logistic regression model predicting staying behaviour from previous points (PrevPoints), previous first-stage states (PrevState), session, drug administrations, and all the interactions between them. Ami, amisulpiride; Nal, naltrexone, Pla, placebo.

Fig. 4 | Effects of amisulpride and naltrexone on the model parameters. The best performing model (M1) is described with three free parameters, ω (the degree of model-based vs model-free value contributions to choice), γ (the degree of devaluation of unencountered spaceships), and η (the inverse temperature in the softmax mapping from values to probabilities). The parameters for both sessions and effects of drug treatments were estimated in one hierarchical model. **a**, Going from session 1 to session 2, amisulpride administration led to higher estimations of ω , and therefore increased model-based relative to model-free control. The difference between sessions of the model-based weights is shown in parameter estimation space (hence the prime). **b**, Difference in the parameter γ . **c**, Difference in the inverse temperature parameter. Lower values mean higher exploration. **d**, Posterior distributions of treatment effects on group-level mean session differences. Ami, amisulpride; Nal, naltrexone, Pla, placebo.

References

1. Dickinson, A. The Development of Behavioural Autonomy. *Philos. Trans. R. Soc. London . Ser. B* (1985).
2. Dolan, J. R. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
3. Balleine, B. W. & O'Doherty, J. P. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology* **35**, 48–69 (2010).
4. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–11 (2005).
5. Patzelt, E. H., Kool, W., Millner, A. J. & Gershman, S. J. Incentives Boost Model-Based Control Across a Range of Severity on Several Psychiatric Constructs. *Biol. Psychiatry* 1–9 (2018). doi:10.1016/j.biopsych.2018.06.018
6. Kool, W., Gershman, S. J. & Cushman, F. A. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychol. Sci.* **28**, 1321–1333 (2017).

7. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20941–20946 (2013).
8. Otto, A. R., Gershman, S. J. & Daw, N. D. The Curse of Planning: Dissecting multiple reinforcement learning systems by taxing the central executive. **24**, (2013).
9. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, J. R. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
10. Kool, W. & Botvinick, M. Mental labour. *Nat. Hum. Behav.* **2**, 899–908 (2018).
11. Everitt, B. J. *et al.* Neural mechanisms underlying the vulnerability to develop compulsive drug-seeking habits and addiction. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 3125–3135 (2008).
12. Voon, V. *et al.* Disorders of compulsivity: a common bias towards learning habits. *Mol. Psychiatry* **20**, 345–352 (2015).
13. Voon, V., Reiter, A. M. F., Sebold, M. & Groman, S. M. Model-Based Control in Dimensional Psychiatry. *Biol. Psychiatry* **82**, 391–400 (2016).
14. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goaldirected control. *Elife* **5**, 1–24 (2016).
15. EMCDDA. European Drug Report 2020: Trends and Developments. (2020).
16. Seth, P., Scholl, L., Rudd, R. A. & Bacon, S. Overdose deaths involving opioids, cocaine, and psychostimulants — United States, 2015–2016. *Am. J. Transplant.* **18**, 1556–1568 (2018).
17. Koob, G. F. & Bloom, F. E. Cellular and molecular mechanisms of drug dependence. *Science* (80-.). (1988). doi:10.1126/science.2903550
18. Chiara, G. Di. Drug addiction as dopamine-dependent associative learning disorder. *Eur. J. Pharmacol.* **375**, 13–30 (1999).
19. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* (80-.). (1997). doi:10.1126/science.275.5306.1593
20. Le Merrer, J., Becker, J. A. J., Befort, K. & Kieffer, B. L. Reward Processing by the Opioid System in the Brain. *Physiol. Rev.* **89**, 1379–1412 (2009).
21. Benjamin, D., Grant, E. R. & Pohorecky, L. A. Naltrexone reverses ethanol-induced dopamine release in the nucleus accumbens in awake, freely moving rats. *Brain Res.* (1993). doi:10.1016/0006-8993(93)90309-B
22. Becker, A., Grecksch, G. & Kraus, J. Rewarding effects of ethanol and cocaine in μ opioid receptor-deficient mice. 296–302 (2002). doi:10.1007/s00210-002-0533-2
23. Gonzalez, J. P. & Brogden, R. N. Naltrexone: A Review of its Pharmacodynamic and Pharmacokinetic Properties and Therapeutic Efficacy in the Management of Opioid Dependence. *Drugs* (1988). doi:10.2165/00003495-198835030-00002
24. Quednow, B. B. & Herdener, M. Human pharmacology for addiction medicine: From evidence to clinical recommendations. in *Progress in Brain Research* (2016). doi:10.1016/bs.pbr.2015.07.017
25. Di Chiara, G. & Bassareo, V. Reward system and addiction: what dopamine does and doesn't do. *Curr. Opin. Pharmacol.* **7**, 69–76 (2007).

26. Wunderlich, K., Smittenaar, P. & Dolan, J. R. Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron* **75**, 418–424 (2012).
27. Kroemer, N. B. *et al.* L-DOPA reduces model-free control of behavior by attenuating the transfer of value to action. *Neuroimage* **186**, 113–125 (2019).
28. Repantis, D., Schlattmann, P., Laisney, O. & Heuser, I. Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacol. Res.* **62**, 187–206 (2010).
29. Cools, R. The costs and benefits of brain dopamine for cognitive control. *Wiley interdisciplinary reviews. Cognitive science* **7**, 317–329 (2016).
30. Volkow, N. D. & Morales, M. The Brain on Drugs: From Reward to Addiction. *Cell* **162**, 712–725 (2015).
31. van Schouwenburg, M., Aarts, E. & Cools, R. Dopaminergic Modulation of Cognitive Control: Distinct Roles for the Prefrontal Cortex and the Basal Ganglia. *Curr. Pharm. Des.* **16**, 2026–2032 (2010).
32. Sawaguchi, T. & Goldman-Rakic, P. D1 dopamine receptors in prefrontal cortex: involvement in working memory. *Science (80-.)*. (1991). doi:10.1126/science.1825731
33. Williams, G. V. & Goldman-Rakic, P. S. Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature* (1995). doi:10.1038/376572a0
34. Goldman-Rakic, P. S. The Cortical Dopamine System: Role in Memory and Cognition. *Adv. Pharmacol.* (1997). doi:10.1016/S1054-3589(08)60846-7
35. Durstewitz, D. & Seamans, J. K. The Dual-State Theory of Prefrontal Cortex Dopamine Function with Relevance to Catechol-O-Methyltransferase Genotypes and Schizophrenia. *Biol. Psychiatry* **64**, 739–749 (2008).
36. Dodds, C. M. *et al.* The dopamine D2 receptor antagonist sulpiride modulates striatal BOLD signal during the manipulation of information in working memory. *Psychopharmacology (Berl)*. **207**, 35–45 (2009).
37. Kool, W., Cushman, F. A. & Gershman, S. J. When Does Model-Based Control Pay Off? *PLoS Comput. Biol.* **12**, e1005090 (2016).
38. Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J. & Lamm, C. Using reinforcement learning models in social neuroscience: Frameworks, pitfalls, and suggestions. *Soc. Cogn. Affect. Neurosci.* (2020).
39. Klaus, J. & Schriefers, H. Measuring verbal working memory capacity: A reading span task for laboratory and web-based use. (2016).
40. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Pers. Soc. Psychol.* (1988). doi:10.1037//0022-3514.54.6.1063
41. Doll, B. B., Bath, K. G., Daw, N. D. & Frank, M. J. Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. *J. Neurosci.* **36**, 1211–1222 (2016).
42. Mannisto, P. T. & Kaakkola, S. Catechol-O-methyltransferase (COMT): Biochemistry, Molecular Biology, Pharmacology, and Clinical Efficacy of the New Selective COMT Inhibitors. *Pharmacol. Rev.* **51**, 593–628 (1999).

43. Trantham-Davidson, H., Neely, L. C., Lavin, A. & Seamans, J. K. Mechanisms underlying differential D1 versus D2 dopamine receptor regulation of inhibition in prefrontal cortex. *J. Neurosci.* **24**, 10652–10659 (2004).
44. Laakso, A. *et al.* The A1 allele of the human D2 dopamine receptor gene is associated with increased activity of striatal L-amino acid decarboxylase in healthy subjects. *Pharmacogenet. Genomics* **15**, 387–391 (2005).
45. Eisenegger, C. *et al.* DAT1 Polymorphism Determines L-DOPA Effects on Learning about Others' Prosociality. *PLoS One* **8**, (2013).
46. Eisenegger, C. *et al.* Role of Dopamine D2 Receptors in Human Reinforcement Learning. *Neuropsychopharmacology* **39**, 2366–2375 (2014).
47. van de Schoot, R. *et al.* Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* **1**, 1 (2021).
48. Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* (1988). doi:10.1080/01621459.1988.10478694
49. van Erp, S., Oberski, D. L. & Mulder, J. Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology* **89**, 31–50 (2019).
50. Deserno, L. *et al.* Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proc. Natl. Acad. Sci.* **112**, 1595–1600 (2015).
51. Voon, V. *et al.* The neurochemical substrates of habitual and goal-directed control. *Transl. Psychiatry* **10**, (2020).
52. Floresco, S. B. & Phillips, A. G. Delay-dependent modulation of memory retrieval by infusion of a dopamine D1 agonist into the rat medial prefrontal cortex. *Behav. Neurosci.* **115**, 934–939 (2001).
53. Frank, M. J. & O'Reilly, R. C. A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behav. Neurosci.* **120**, 497–517 (2006).
54. Seamans, J. K. & Yang, C. R. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Prog. Neurobiol.* **74**, 1–58 (2004).
55. O'Donnell, P. & Grace, A. A. Tonic D2-mediated attenuation of cortical excitation in nucleus accumbens neurons recorded in vitro. *Brain Res.* **634**, 105–112 (1994).
56. Mehta, M. A., Manes, F. F., Magnolfi, G., Sahakian, B. J. & Robbins, T. W. Impaired set-shifting and dissociable effects on tests of spatial working memory following the dopamine D2 receptor antagonist sulpiride in human volunteers. *Psychopharmacology (Berl)*. **176**, 331–342 (2004).
57. Kahnt, T., Weber, S. C., Haker, H., Robbins, T. W. & Tobler, P. N. Dopamine D2-receptor blockade enhances decoding of prefrontal signals in humans. *J. Neurosci.* **35**, 4104–4111 (2015).
58. Naef, M. *et al.* Effects of dopamine D2/D3 receptor antagonism on human planning and spatial working memory. *Transl. Psychiatry* **7**, e1107 (2017).
59. Mehta, M. A., Sahakian, B. J., McKenna, P. J. & Robbins, T. W. Systemic sulpiride in young adult volunteers simulates the profile of cognitive deficits in Parkinson's disease.

Psychopharmacology (Berl). **146**, 162–174 (1999).

60. Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D. & Dolan, J. R. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron* **80**, 914–919 (2013).
61. Westbrook, A., Frank, M. J. & Cools, R. A mosaic of cost–benefit control over cortico-striatal circuitry. *Trends Cogn. Sci.* **25**, 710–721 (2021).
62. Westbrook, A. *et al.* Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science (80-.)*. **367**, 1362–1366 (2020).
63. Ford, C. P. The role of D2-autoreceptors in regulating dopamine neuron activity and transmission. *Neuroscience* (2014). doi:10.1016/j.neuroscience.2014.01.025
64. Jocham, G., Klein, T. A. & Ullsperger, M. Differential Modulation of Reinforcement Learning by D2 Dopamine and NMDA Glutamate Receptor Antagonism. *J. Neurosci.* **34**, 13151–13162 (2014).
65. Pessiglione, M., Seymour, B., Flandin, G., Dolan, J. R. & Frith, C. D. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* (2006). doi:10.1038/nature05051
66. Weber, S. C. *et al.* Dopamine D2/3- and μ -opioid receptor antagonists reduce cue-induced responding and reward impulsivity in humans. *Transl. Psychiatry* **6**, e850 (2016).
67. Korb, S. *et al.* Dopaminergic and opioidergic regulation during anticipation and consumption of social and nonsocial rewards. *Elife* **9**, 1–22 (2020).
68. Adams, R. A. *et al.* Variability in Action Selection Relates to Striatal Dopamine 2/3 Receptor Availability in Humans: A PET Neuroimaging Study Using Reinforcement Learning and Active Inference Models. *Cereb. Cortex* **30**, 3573–3589 (2020).
69. Lee, M. C. *et al.* Duration of occupancy of opiate receptors by naltrexone. *J. Nucl. Med.* (1988).
70. Soutschek, A., Weber, S. C., Kahnt, T., Quednow, B. B. & Tobler, P. N. Opioid antagonism modulates wanting-related frontostriatal connectivity. *Elife* **10**, 1–17 (2021).
71. Delamater, A. R., Sclafani, A. & Bodnar, R. J. Pharmacology of Sucrose-Reinforced Place-Preference Conditioning: Effects of Naltrexone. *Pharmacol. Biochem. Behav.* **65**, 697–704 (2000).
72. Ågmo, A., Galvan, A. & Talamantes, B. Reward and reinforcement produced by drinking sucrose: Two processes that may depend on different neurotransmitters. *Pharmacol. Biochem. Behav.* **52**, 403–414 (1995).
73. Wassum, K. M., Cely, I. C., Maidment, N. T. & Balleine, B. W. Disruption of endogenous opioid activity during instrumental learning enhances habit acquisition. *Neuroscience* **163**, 770–780 (2009).
74. van Steenbergen, H., Eikemo, M. & Leknes, S. The role of the opioid system in decision making and cognitive control: A review. *Cogn. Affect. Behav. Neurosci.* **19**, 435–458 (2019).
75. Feher da Silva, C. & Hare, T. A. Humans primarily use model-based inference in the two-stage task. *Nat. Hum. Behav.* **4**, 1053–1066 (2020).
76. Daw, N. D. Are we of two minds? *Nat. Neurosci.* **21**, 1497–1499 (2018).

77. Melnikoff, D. E. & Bargh, J. A. The Mythical Number Two. *Trends Cogn. Sci.* (2018). doi:10.1016/j.tics.2018.02.001
78. World Medical Association Declaration of Helsinki. *JAMA* **310**, 2191 (2013).
79. Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning. *Psychol. Sci.* **27**, 848–858 (2016).
80. Kruschke, J. K. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition.* Academic Press (2014). doi:10.1016/B978-0-12-405888-0.09999-2
81. Bürkner, P.-C. brms : An R Package for Bayesian Multilevel Models Using Stan . *J. Stat. Softw.* (2017). doi:10.18637/jss.v080.i01
82. Carpenter, B. et al. Stan: A probabilistic programming language. *J. Stat. Softw.* (2017). doi:10.18637/jss.v076.i01
83. Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction. *UCL, Computer Sci. Dep. Reinf. Learn. Lect.* 1054 (2017). doi:10.1109/TNN.1998.712192
84. Ahn, W.-Y., Haines, N. & Zhang, L. Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Comput. Psychiatry* (2017). doi:10.1162/cpsy_a_00002
85. McElreath, R. *Statistical rethinking: A bayesian course with examples in R and stan. Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2018). doi:10.1201/9781315372495
86. Lengersdorff, L., Wagner, I. & Lamm, C. When implicit prosociality trumps selfishness: the neural valuation system underpins more optimal choices when learning to avoid harm to others than to oneself. *preprint* (2020). doi:https://doi.org/10.31234/osf.io/q6psx
87. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* (1992). doi:10.1214/ss/1177011136
88. Nalborczyk, L., Batailler, C., Loevenbruck, H., Vilain, A. & Bürkner, P.-C. An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *J. Speech, Lang. Hear. Res.* **62**, 1225–1242 (2019).
89. Hedges, L. V. Effect sizes in cluster-randomized designs. (2007).
90. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* (2017). doi:10.1007/s11222-016-9696-4
91. Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Anal.* (2018). doi:10.1214/17-ba1091