Open access • Posted Content • DOI:10.1101/2020.02.20.958025

Noradrenaline modulates tabula-rasa exploration — Source link

Magda Dubois, Johanna Habicht, Jochen Michely, Rani Moran ...+2 more authors

Published on: 21 Feb 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Human complex exploration strategies are extended via noradrenaline-modulated heuristics

Dubois M[1,2], Habicht J[1,2], Michely J[1,2], Moran R[1,2], Dolan RJ[1,2] & Hauser TU[1,2]

[1]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom.
[2]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom.

Corresponding author

Tobias U. Hauser
Max Planck UCL Centre for Computational Psychiatry and Ageing Research
University College London
10-12 Russell Square
London WC1B 5EH
United Kingdom
Phone: +44 / 207 679 5264
Email: t.hauser@ucl.ac.uk

Number of pages: 67
Number of Figures: 5
Number of Tables: 0
Abstract: 123 words
Introduction: 1031 words
Discussion: 2393 words

Data and materials availability: Data and code will be provided upon acceptance.

Abstract

An exploration-exploitation trade-off, the arbitration between sampling a lesser-known against a known rich option, is thought to be solved using computationally demanding exploration algorithms. Given known limitations in human cognitive resources, we hypothesised the presence of additional cheaper strategies. We examined for such heuristics in choice behaviour where we show this involves a value-free random exploration, that ignores all prior knowledge, and a novelty exploration that targets novel options alone. In a double-blind, placebo-controlled drug study, assessing contributions of dopamine (400mg amisulpride) and noradrenaline (40mg propranolol), we show that value-free random exploration is attenuated under the influence of propranolol, but not under amisulpride. Our findings demonstrate that humans deploy distinct computationally cheap exploration strategies and where value-free random exploration is under noradrenergic control.

Introduction

Chocolate, Toblerone, spinach or hibiscus ice-cream? Do you go for the flavour you like the most (chocolate), or another one? In such an exploration-exploitation dilemma, you need to decide whether to go for the option with the highest known subjective value (exploitation) or opt instead for less known or valued options (exploration) so as to not miss out on possibly even higher rewards. In the latter case, you can opt to either chose an option that you have previously enjoyed (Toblerone), an option you are curious about because you do not know what to expect (hibiscus), or even an option that you have disliked in the past (spinach). Depending on your exploration strategy, you may end up with a highly disappointing ice cream encounter, or a life-changing gustatory epiphany.

A common approach to the study of complex decision making, for example an exploration-exploitation trade-off, is to take computational algorithms developed in the field of artificial intelligence and test whether key signatures of these are evident in human behaviour. This approach has revealed humans use strategies that reflect an implementation of computationally demanding exploration algorithms (*1*, *2*). One such strategy, directed exploration, involves awarding an 'information bonus' to choice options, a bonus that scales with uncertainty. This is captured in algorithms such as the Upper Confidence Bound (UCB) (*3*, *4*) and leads to an exploration of choice options the agent knowns little about (*1*, *5*) (e.g. the hibiscus ice-cream). An alternative strategy, sometimes termed 'random' exploration, is to induce stochasticity after value computations in the decision process. This can be realised using a fixed parameter as a source of stochasticity, such as a softmax temperature parameter (*6*, *7*), which can be combined with the UCB algorithm (*1*). Alternatively, one can use a dynamic source of stochasticity, such as in Thompson sampling (*8*), where stochasticity adapts to an uncertainty about choice options. This exploration is essentially a more sophisticated, uncertainty-driven, version of a softmax. By accounting for stochasticity when comparing choice options' expected values, in effect choosing based on both uncertainty and value, these exploration strategies increase the likelihood of choosing 'good' options that are only slightly less valuable than the best (e.g. the Toblerone ice-cream if you are a chocolate lover).

The above processes are computationally demanding, especially when facing real-life multiple-alternative decision problems (6, 9, 10). Human cognitive resources are constrained by capacity limitations (11), metabolic consumption (12), but also because of resource allocation to parallel tasks (e.g. (*13*, *14*)). This directly relates to an agents' motivation to perform a given task (*11*, *15*, *16*), as increasing an information demand in one process automatically reduces its availability for others (12). In real-world highly dynamic environments, this arbitration is critical as humans need to maintain resources for alternative opportunities (i.e. flexibility; (*11*, *17*, *18*)). This accords with previous studies showing humans are demand-avoidant (*17*, *19*) and suggests that exploration computations tend to be minimised. Here, we examine the explanatory power of two additional computationally less costly forms of exploration, namely value-free random exploration and novelty exploration.

Computationally, the least resource demanding way to explore is to ignore all prior information and to choose entirely randomly, de facto assigning the same probability to all options. Such 'value-free' random exploration, as opposed to the two previously considered 'value-based' random explorations (for simulations comparing their effects cf. Figure 1 – Figure supplement 2) that add stochasticity during choice value computation, forgoes any costly computation (i.e. value mean and uncertainty), known as an ϵ-greedy algorithmic strategy in reinforcement learning (*20*).

Computational efficiency, however, comes at the cost of sub-optimality due to occasional selection of options of low expected value (e.g. the repulsive spinach ice cream).

Despite its sub-optimality, value-free random exploration has neurobiological plausibility. Of relevance in this context is a view that exploration strategies depend on dissociable neural mechanisms (*21*). Influences from noradrenaline and dopamine are plausible candidates in this regard based on prior evidence (*9*, *22*). Amongst other roles (such as memory (*23*), or energisation of behaviour (*24*, *25*)), the neuromodulator noradrenaline has been ascribed a function of indexing uncertainty (*26–28*) or as acting as a 'reset button' that interrupts ongoing information processing (*29–31*). Prior experimental work in rats shows boosting noradrenaline leads to more tabula-rasa-like random behaviour (*32*), while pharmacological manipulations in monkeys indicates reducing noradrenergic activity increases choice consistency (*33*).

In human pharmacological studies, interpreting the specific function of noradrenaline on exploration strategies is problematic as many drugs, such as atomoxetine (e.g. (*34*)), impact multiple neurotransmitter systems. Here, to avoid this issue, we chose the highly specific β-adrenoceptor antagonist propranolol, which has only minimal impact on other neurotransmitter systems (*35–37*). Using this neuromodulator, we examine whether signatures of value-free random exploration are impacted by administration of propranolol.

An alternative computationally efficient exploration heuristic to random exploration is to simply choose an option not encountered previously, which we term novelty exploration. Humans often show novelty seeking (*38–41*), and this strategy can be used in exploration as implemented by a low-cost version of the UCB algorithm. Here a novelty bonus (*42*) is added if a choice option has not been seen previously (i.e. it does not have to rely on precise uncertainty estimates). The neuromodulator dopamine is implicated not only in exploration in general (*43*), but also in signalling such types of novelty bonuses, where evidence indicates a role in processing and exploring novel and salient states (*39*, *44–47*). Although pharmacological dopaminergic studies in humans have demonstrated effects on exploration as a whole (*48*), they have not identified specific exploration strategies. Here, we used the highly specific D2/D3 antagonist, amisulpride, to disentangle the specific role of dopamine and noradrenaline on different exploration strategies.

Thus, in the current study, we examine the contributions of value-free random exploration and novelty exploration in human choice behaviour. We developed a novel exploration task combined with computational modeling to probe the contributions of noradrenaline and dopamine. Under double-blind, placebo-controlled, conditions we tested the impact of two antagonists with a high affinity and specificity for either dopamine (amisulpride) or noradrenaline (propranolol). Our results provide evidence that both exploration heuristics supplement computationally more demanding exploration strategies, and that value-free random exploration is particularly sensitive to noradrenergic modulation.

Results

Probing the contributions of heuristic exploration strategies

We developed a novel multi-round three-armed bandit task (Figure 1; bandits depicted as trees), enabling us to assess the contributions of value-free random exploration and novelty exploration in addition to Thompson sampling and UCB (combined with a softmax). In particular, we exploited the fact that both heuristic strategies make specific predictions about choice patterns. The novelty exploration assigns a 'novelty bonus' only to bandits for which subjects have no prior information, but not to other bandits. This can be seen as a low-resolution version of UCB, which assigns a bonus to all choice options proportionally to how informative they are, in effect a graded bonus which scales to each bandits' uncertainty. Thus, to capture this heuristic, we manipulated the amount of prior information with bandits carrying only little information (i.e. 1 vs 3 initial samples) or no information (0 initial samples). A high novelty exploration predicts a higher frequency of selecting the novel option (Figure 1f). This is in contrast to high exploration using other strategies which does not predict such a strong effect on the novel option (cf. Figure 1 - Figure supplement 5).

Value-free random exploration, captured here by ϵ-greedy, predicts that all prior information is discarded entirely and that there is equal probability attached to all choice options. This strategy is distinct from other exploration strategies as it is likely to choose bandits known to be substantially worse than the other bandits. Thus, a high value-free random exploration predicts a higher frequency of selecting the low-value option (Figure 1e), whereas high exploration using other strategies does not predict such effect (cf. Figure 1 - Figure supplement 3). A second prediction is that choice consistency, across repeated trials, is substantially affected by value-free random exploration. Given that value-free random exploration splits its choice probability equally (i.e. 33.3% of choosing any bandit out of the three displayed), an increase in such exploration predicts a lower likelihood of choosing the same bandit again, even under identical choice options (Figure 1e). This contrasts to other strategies that make consistent exploration predictions (e.g. UCB would consistently explore the choice option that carries a high information bonus; Figure 1 - Figure supplement 4).

We generated bandits from four different generative processes (Figure 1c) with distinct sample means (but a fixed sampling variance) and number of initial samples (i.e. samples shown at the beginning of a trial for this specific bandit). Subjects were exposed to these bandits before making their first draw. The 'certain-standard bandit' and the (less certain) 'standard bandit' were bandits with comparable means but varying levels of uncertainty, providing either three or one initial samples (depicted as apples; similar to the horizon task (*7*)). The 'low-value bandit' was a bandit with one initial sample from a substantially lower generative mean, thus appealing to a value-free random exploration strategy alone. The last bandit, with a mean comparable with that of the standard bandits, was a 'novel bandit' for which no initial sample was shown, primarily appealing to a novelty exploration strategy (cf. Materials and Methods for a full description of bandit generative processes). To assess choice consistency, all trials were repeated once. In the pilot experiments (data not shown), we noted some exploration strategies tended to overshadow other strategies. To effectively assess all exploration strategies, we opted to present only three of the four different bandit types on each trial, as different bandit triples allow different explorations to manifest. Lastly, to assess whether subjects' behaviour captured exploration, we manipulated the degree to which subjects could interact with the same bandits. Similar to previous studies (*7*), subjects could perform either one draw, encouraging exploitation (short horizon condition) or six draws encouraging more substantial explorative behaviour (long horizon condition) (*7*, *34*).

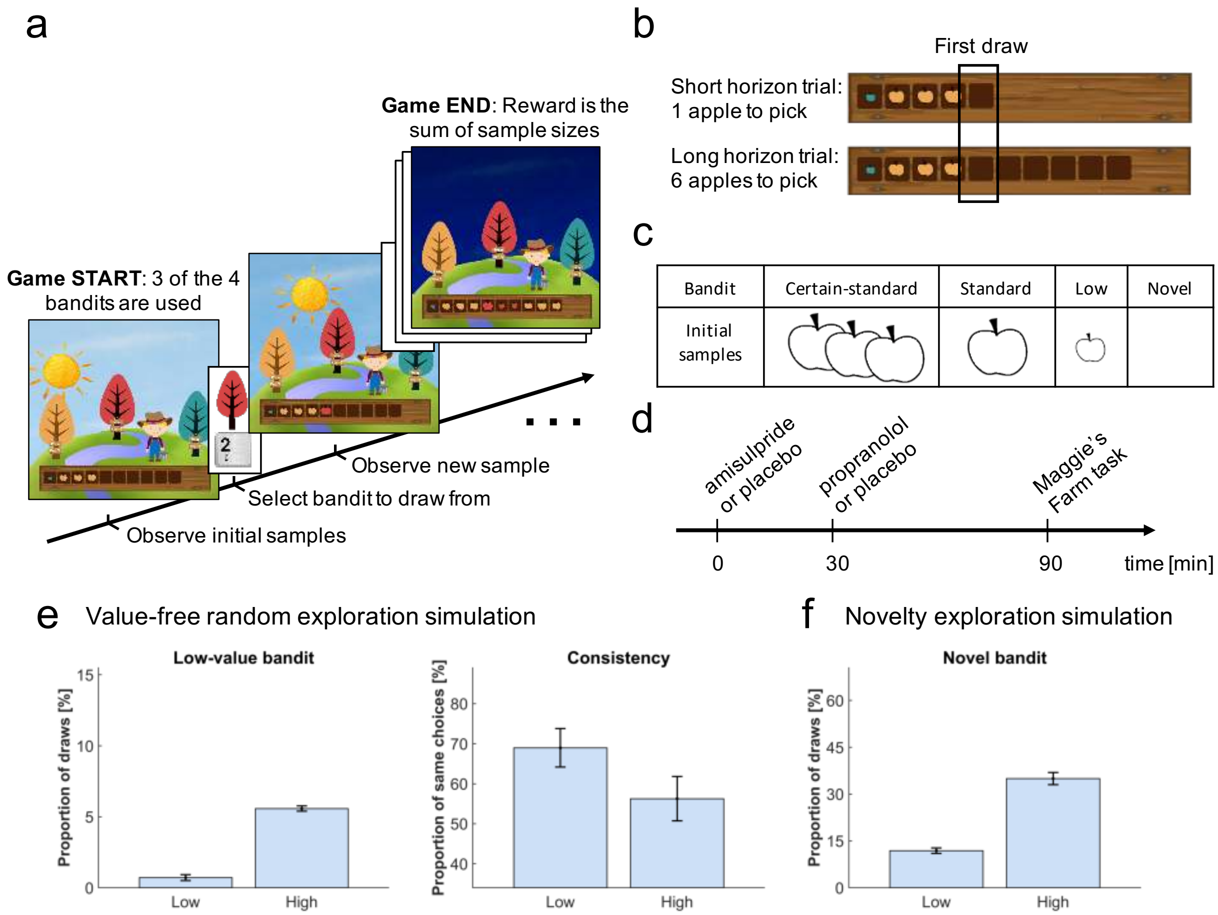


Figure 1. Study design. In the Maggie's farm task, subjects had to choose from three bandits (depicted as trees) to maximise an outcome (sum of reward). The rewards (apple size) of each bandit followed a normal distribution with a fixed sampling variance. (a) At the beginning of each trial, subjects were provided with some initial samples on the wooden crate at the bottom of the screen and had to select which bandit they wanted to sample from next. (b) Depending the condition, they could either perform one draw (short horizon) or six draws (long horizon). The empty spaces on the wooden crate (and the suns' position) indicated how many draws they had left. The first draw in both conditions was the main focus of the analysis. (c) In each trial, three bandits were displayed, selected from four possible bandits, with different generative processes that varied in terms of their sample mean and number of initial samples (i.e. samples shown at the beginning of a trial). The 'certain-standard bandit' and the 'standard bandit' had comparable means but different levels of uncertainty about their expected mean: they provided three and one initial sample respectively; the 'low-value bandit' had a low mean and displayed one initial sample; the 'novel bandit' did not show any initial sample and its mean was comparable with that of the standard bandits. (d) Prior to the task, subjects were administered different drugs: 400mg amisulpride that blocks dopaminergic D2/D3 receptors, 40mg propranolol to block noradrenergic β-receptors, and inert substances for the placebo group. Different administration times were chosen to comply with the different drug pharmacokinetics (placebo matching the other groups' administration schedule). (e) Simulating value-free random behaviour with a low vs high model parameter (ϵ) in this task shows that in a high regime, agents choose the low-value bandit more often (left panel; mean ± SD) and are less consistent in their choices when facing identical choice

options (right panel). (f) Novelty exploration exclusively promotes choosing choice options for which subjects have no prior information, captured by the 'novel bandit' in our task. For details about simulations cf. Materials and Methods. For details about the task display cf. Figure 1 – Figure supplement 1. For simulations of different exploration strategies and their impact of different bandits cf. Figure 1 – Figure supplement 2-5.

Testing the role of catecholamines noradrenaline and dopamine

In a double-blind, placebo-controlled, between-subjects, study design we assigned subjects (N=60) randomly to one of three experimental groups: amisulpride, propranolol or placebo. The first group received 40mg of the β-adrenoceptor antagonist propranolol to alter noradrenaline function, while the second group was administered 400mg of the D2/D3 antagonist amisulpride that alters dopamine function. Because of different pharmacokinetic properties, these drugs were administered at different times (Figure 1d) and compared to a placebo group that received a placebo at both drug times to match the corresponding antagonists' time. One subject (amisulpride group) was excluded from the analysis due to a lack of engagement with the task. Reported findings were corrected for IQ and mood, as drug groups differed marginally in those measures (cf. Appendix 2 Table 1), by adding WASI (*49*) and PANAS (*50*) negative scores as covariates in each ANOVA. Similar results were obtained in an analysis that corrected for physiological effects as from the analysis without covariates (cf. Appendix 1).

Increased exploration when information can subsequently be exploited

Our task embodied two decision-horizon conditions, a short and a long. To assess whether subjects explored more in a long horizon condition, in which additional information can inform later choices, we examined which bandit subjects chose in their first draw (in accordance with the horizon task (*7*)), irrespective of their drug group. A marker of exploration here is evident if subjects chose bandits with lower expected values, computed as the mean value of their initial samples shown (trials where the novel bandit was chosen were excluded). As expected, subjects chose bandits with a lower expected value in the long compared to the short horizon (repeated-measures ANOVA for the expected value: F(1, 56)=19.457, p<.001, η2=.258; Figure 2a). To confirm that this was a consequence of increased exploration, we analysed the proportion of how often the high-value option was chosen (i.e. the bandit with the highest expected reward based on its initial samples) and we found that subjects (especially those with higher IQ) sampled from it more in the short compared to the long horizon, (WASI-by-horizon interaction: F(1,54)=13.304, p=.001, η2=.198; horizon main effect: F(1, 54)=3.909, p=.053, η2=.068; Figure 3a), confirming a reduction in exploitation when this information could be subsequently used. Interestingly, this frequency seemed to be marginally higher in the amisulpride group, suggesting an overall higher tendency to exploitation following dopamine blockade (cf. Appendix 1). This horizon-specific behaviour resulted in a lower reward on the 1st sample in the long compared to the short horizon (F(1, 56)=23.922, p<.001, η2=.299; Figure 2c). When we tested whether subjects were more likely to choose options they knew less about (computed as the mean number of initial samples shown), we found that subjects chose less known (i.e. more informative) bandits more often in the long horizon compared to the short horizon (F(1, 56)=58.78, p<.001, η2=.512; Figure 2b).

Next, to evaluate whether subjects used the additional information beneficially in the long horizon condition, we compared the average reward (across six draws) obtained in the long compared to short horizon (one draw). We found that the average reward was higher in the long horizon (F(1, 56)=103.759, p<.001, η2=.649; Figure 2c), indicating that subjects tended to choose less optimal bandits at first but subsequently learnt to appropriately exploit the harvested information to guide choices of better bandits in the long run. Additionally, when looking specifically at the long horizon condition, we found that subjects earned more when their first draw was explorative versus exploitative (Figure 2 - Figure supplement 1c-d; cf. Appendix 2 for details).

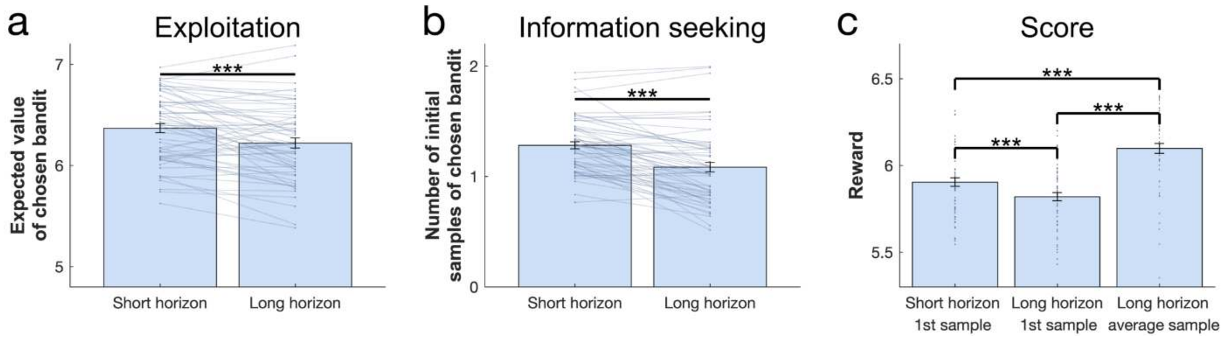


Figure 2. Benefits of exploration. To investigate the effect of information on performance we collapsed subjects over all three treatment groups. (a) The expected value (average of its initial samples) of the first chosen bandit as a function of horizon. Subjects chose bandits with a lower expected value (i.e. they explored more) in the long horizon compared to the short horizon. (b) The mean number of samples for the first chosen bandit as a function of horizon. Subjects chose less known (i.e. more informative) bandits more in the long compared to the short horizon. (c) The first draw in the long horizon led to a lower reward than the first draw in the short horizon, indicating that subjects sacrificed larger initial outcomes for the benefit of more information. This additional information helped making better decisions in the long run, leading to a higher earning over all draws in the long horizon. For values and statistics cf. Appendix 2 Table 3. For response times and details about all long horizons' samples cf. Figure 2 – Figure supplement 1. *** =p<.001. Data are shown as mean ± SEM and each dot/line represent a subject.

Subjects demonstrate value-free random behaviour

Value-free random exploration (analogue to ϵ-greedy) predicts that ϵ % of the time each option will have an equal probability of being chosen. In such a regime (compared to more complex strategies that would favour options with a higher expected value with a similar uncertainty), the probability of choosing bandits with a low expected value (here the low-value bandit; Fig. 1e) will be higher (cf. Figure 1 – Figure supplement 3). We investigated whether the frequency of picking the low-value bandit was increased in the long horizon condition across all subjects (i.e. when exploration is useful), and we found a significant main effect of horizon (F(1, 54)=4.069, p=.049, η2=.07; Figure 3b). This demonstrates that value-free random exploration is utilised more when exploration is beneficial.

Value-free random behaviour is modulated by noradrenaline function

When we tested whether value-free random exploration was sensitive to neuromodulatory influences, we found a difference in how often drug groups sampled from the low-value option (drug main effect: F(2, 54)=7.003, p=.002, η2=.206; drug-by-horizon interaction: F(2, 54)=2.154, p=.126, η2=.074; Figure 3b). This was driven by the propranolol group choosing the low-value option significantly less often than the other two groups (placebo vs propranolol: t(40)=2.923, p=.005, d=.654; amisulpride vs propranolol: t(38)=2.171, p=.034, d=.496) with no difference between amisulpride and placebo: (t(38)=-0.587, p=.559, d=.133). These findings demonstrate that a key feature of value-free random exploration, the frequency of choosing low-value bandits, is sensitive to influences from noradrenaline.

To further examine drug effects on value-free random exploration, we assessed a second prediction, namely choice consistency. Because value-free random exploration ignores all prior information and chooses randomly, it should result in a decreased choice consistency when presented identical choice options (cf. Figure 1 – Figure supplement 2 & 4, compared to more complex strategies which are always biased towards the rewarding or the information providing bandit for example). To this end, each trial was duplicated in our task, allowing us to compute the consistency as the percentage of time subjects sampled from an identical bandit when facing the exact same choice options. In line with the above analysis, we found a difference in consistency by which drug groups sampled from different option (drug main effect: F(2, 54)=7.154, p=.002, η2=.209; horizon main effect: F(1, 54)=1.333, p=.253, η2=.024; drug-by-horizon interaction: F(2, 54)=3.352, p=.042, η2=.11; Figure 3c), driven by the fact that the propranolol group chose significantly more consistently than the other two groups (pairwise comparisons: placebo vs propranolol: t(40)=-3.525, p=.001, d=.788; amisulpride vs placebo: t(38)=1.107, p=.272, d=.251; amisulpride vs propranolol: t(38)=-2.267, p=.026, d=.514). Please see Appendix 1 for further discussion and analysis of the drug-by-horizon interaction. Taken together, these results indicate that value-free random exploration depends critically on noradrenaline functioning, such that an attenuation of noradrenaline leads to a reduction in value-free random exploration.

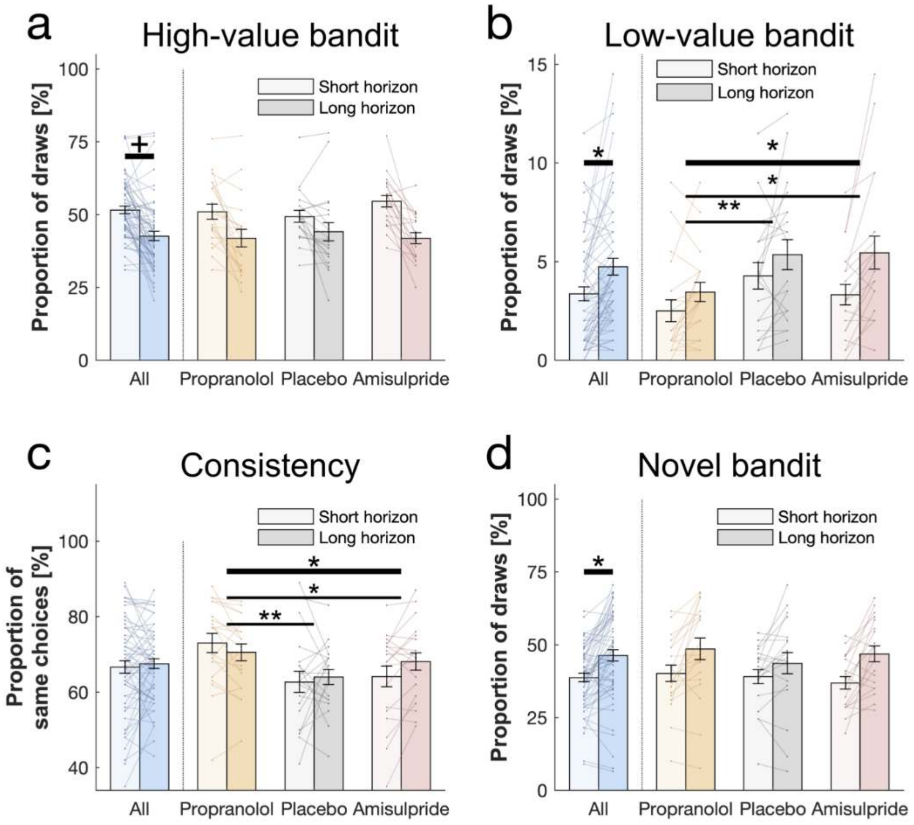


Figure 3. Behavioural horizon and drug effects. Choice patterns in the first draw for each horizon and drug group (propranolol, placebo and amisulpride). (a) Subjects sampled from the high-value bandit (i.e. bandit with the highest average reward of initial samples) more in the short horizon compared to the long horizon indicating reduced exploitation. (b) Subjects sampled from the low-value bandit more in the long horizon compared to the short horizon indicating value-free random exploration, but subjects in the propranolol group sampled less from it overall, and (c) were more consistent in their choices overall, indicating that noradrenaline blockade reduces value-free random exploration. (d) Subjects sampled from the novel bandit more in the long horizon compared to the short horizon indicating novelty exploration. Please note that some horizon effects were modulated by subjects' intellectual abilities when additionally controlling for them (cf.

Appendix 2 Table 4). Horizontal bars represent rm-ANOVA (thick) and pairwise comparisons (thin). † =p<.07, * =p<.05, ** =p<.01. Data are shown as mean ± SEM and each line represent one subject. For values and statistics cf. Appendix 2 Table 4. For response times and frequencies specific to the displayed bandits cf. Figure 3 – Figure supplement 1-2.

Novelty exploration is unaffected by catecholaminergic drugs

Next, we examined whether subjects show evidence for novelty exploration by choosing the novel bandit for which there was no prior information (i.e. no initial samples), as predicted by model simulations (Figure 1f). We found a significant main effect of horizon (F(1, 54)=5.593, p=.022, η2=.094; WASI-by-horizon interaction: F(1, 54) =13.897, p<.001, η2=.205; Figure 3d) indicating that subjects explored the novel bandit significantly more often in the long horizon condition, and this was particularly strong for subjects with a higher IQ. We next assessed whether novelty exploration was sensitive to our drug manipulation, but found no drug effects on the novel bandit (F(2, 54)=1.498, p=.233, η2=.053; drug-by-horizon interaction: F(2, 54)=.542, p=.584, η2=.02; Figure 3d). Thus, there was no evidence that an attenuation of dopamine or noradrenaline function impact novelty exploration in this task.

Subjects combine computationally demanding strategies and exploration heuristics

To examine the contributions of different exploration strategies to choice behaviour, we fitted a set of computational models to subjects' behaviour, building on models developed in previous studies (*1*). In particular, we compared models incorporating UCB, Thompson sampling, an ϵ-greedy algorithm and the novelty bonus (cf. Materials and Methods). Essentially, each model makes different exploration predictions. In the Thompson model, Thompson sampling (*8*, *51*) leads to an uncertainty-driven value-based random exploration, where both expected value and uncertainty contribute to choice. In this model higher uncertainty leads to more exploration such that instead of selecting a bandit with the highest mean, bandits are chosen relative to how often a random sample would yield the highest outcome, thus accounting for uncertainty (*2*). The UCB model (*3*, *4*), capturing directed exploration, predicts that each bandit is chosen according to a mixture of expected value and an additional expected information gain (*2*). This is realised by adding a bonus to the expected value of each option, proportional to how informative it would be to select this option (i.e. the higher the uncertainty in the options' value, the higher the information gain). This computation is then passed through a softmax decision model, capturing value-based random exploration. Novelty exploration is a simplified version of the information bonus in the UCB algorithm, which only applies to entirely novel options. It defines the intrinsic value of selecting a bandit about which nothing is known, and thus saves demanding computations of uncertainty for each bandit. Lastly, the value-free random ϵ-greedy algorithm selects any bandit ϵ % of the time, irrespective of the prior information of this bandit. For additional models cf. Appendix 1.

We used cross-validation for model selection (Figure 4a) by comparing the likelihood of held-out data across different models, an approach that adequately arbitrates between model accuracy and complexity. The winning model encompasses uncertainty-driven value-based random exploration (Thompson sampling) with value-free random exploration (ϵ-greedy parameter) and novelty exploration (novelty bonus parameter η). The winning model predicted held-out data with a 55.25% accuracy (SD=8.36%; chance level =33.33%). Similarly to previous studies (*1*), the hybrid model combining UCB and Thompson sampling explained the data better than each of those processes alone, but this was no longer the case when accounting for novelty and value-free random exploration (Figure 4a). The winning model further revealed that all parameter estimates could be accurately recovered (Figure 4b; Figure 4 – Figure supplement 3). Interestingly, although the 2nd and 3rd place models made different prediction about the complex exploration strategy, using a directed exploration with value-based random exploration (UCB) or a combination of complex strategies (hybrid) respectively, they share the characteristic of benefitting from value-free random and novelty exploration. This highlights that subjects used a mixture of computationally demanding and heuristic exploration strategies.

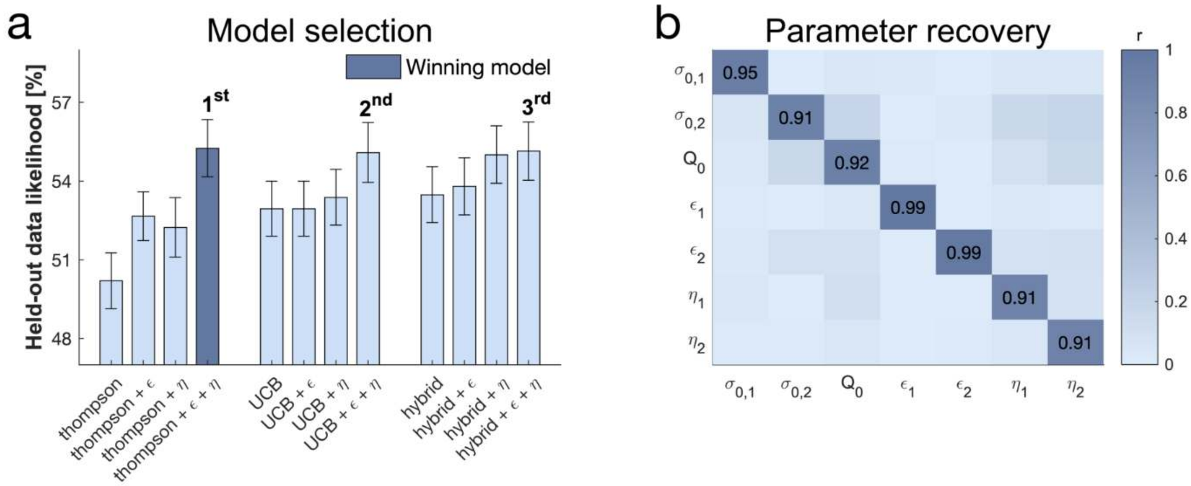


Figure 4. Subjects use a mixture of exploration strategies. (a) A 10-fold cross-validation of the likelihood of held-out data was used for model selection (chance level =33.3%; for model selection at the individual level cf. Figure 4 – Figure supplement 1). The Thompson model with both the ϵ-greedy parameter and the novelty bonus η best predicted held-out data (b) Model simulation with 4^7 simulations predicted good recoverability of model parameters (for correlations between behaviour and model parameters cf. Figure 4 – Figure supplement 2); σ_0 is the prior variance and Q_0 is the prior mean (for parameter recovery correlation plots cf. Figure 4 – Figure supplement 3). 1 stands for short horizon-, and 2 for long horizon-specific parameters. For values and parameter details cf. Appendix 2 Table 5.

Noradrenaline controls value-free random exploration

To more formally compare the impact of catecholaminergic drugs on different exploration strategies, we assessed the free parameters of the winning model between drug groups (Figure 5, cf. Appendix 2 Table 6 for exact values). First, we examined the ϵ-greedy parameter that captures the contribution of value-free random exploration to choice behaviour. We assessed how this value-free random exploration differed between drug groups. A significant drug main effect (drug main effect: F(2, 54)=6.722, p=.002, η2=.199; drug-by-horizon interaction: F(2, 54)=1.305, p=.28, η2=.046; Figure 5a) demonstrates that the drug groups differ in how strongly they deploy this exploration strategy. Post-hoc analysis revealed that subjects with reduced noradrenaline functioning had the lowest values of ϵ (pairwise comparisons: placebo vs propranolol: t(40)=3.177, p=.002, d=.71; amisulpride vs propranolol: t(38)=2.723, p=.009, d=.626) with no significant difference between amisulpride vs placebo: (t(38)=.251, p=.802, d=.057). Critically, the effect on ϵ was also significant when the complex exploration strategy was a directed exploration with value-based random exploration (2nd place model) and, marginally significant, when it was a combination of the above (3rd place model; cf. Appendix 1).

The ϵ-greedy parameter was also closely linked to the above behavioural metrics (correlation between the ϵ-greedy parameter with draws from the low-value bandit: $R_{Pearson}$=.828, p<.001; and with choice consistency: $R_{Pearson}$=-.596, p<.001; Figure 4 – Figure supplement 2), and showed a similar horizon effect (horizon main effect: F(1, 54)=1.968, p=.166, η2=.035; WASI-by-horizon interaction: F(1, 54)=6.08, p=.017, η2=.101; Figure 5a). Our findings thus accord with the model-free analyses and demonstrate that noradrenaline blockade reduces value-free random exploration.

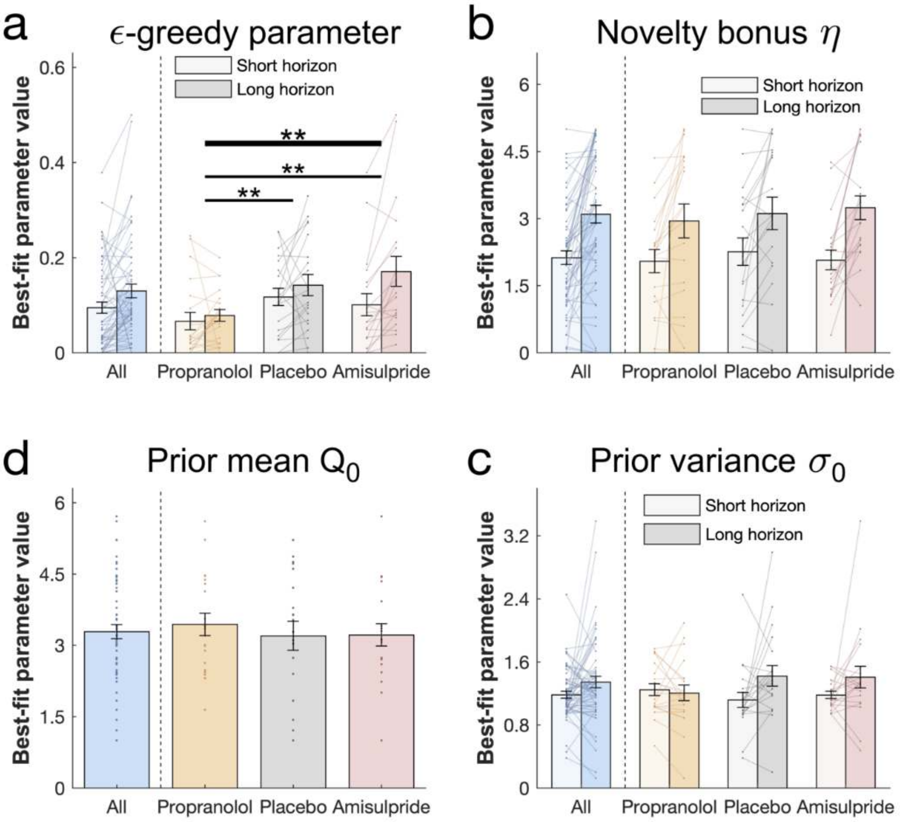


Figure 5. Drug effects on model parameters. The winning model's parameters were fitted to each subject's first draw (for model simulations cf. Figure 5 – Figure supplement 1). (a) Subjects had higher values of ϵ (value-free random exploration) in the long compared to the short horizon. Notably, subjects in the propranolol group had lower values of ϵ overall, indicating that attenuation of noradrenaline functioning reduces value-free random exploration. Subjects from all groups (b) assigned a similar value to novelty, captured by the novelty bonus η, which was higher (more novelty exploration) in the long compared to the short horizon. (c) The groups had similar beliefs Q_0 about a bandits' mean before seeing any initial samples and (d) were similarly uncertain σ_0 about it (for gender effects cf. Figure 5 – Figure supplement 2). Please note that some horizon effects were modulated by subjects' intellectual abilities when additionally controlling for them (cf. Appendix 2 Table 6). ** =p<.01. Data are shown as mean ± SEM and each dot/line represent one subject. For parameter values and statistics cf. Appendix 2 Table 6.

No drug effects on other parameters

The novelty bonus η captures the intrinsic reward of selecting a novel option. In line with the model-free behavioural findings, there was no difference between drug groups in terms of this effect (F(2, 54)=.249, p=.78, η^2=.009; drug-by-horizon interaction: F(2, 54)=.03, p=.971, η^2=.001). There was also a close alignment between model-based and model-agnostic analyses (correlation between the novelty bonus η with draws from the novel bandit: $R_{Pearson}$=.683, p<.001; Figure 4 – Figure supplement 2), and we found a similarly increased novelty bonus effect in the long horizon in subjects with a higher IQ (WASI-by-horizon interaction: F(1, 54) =8.416, p=.005, η^2=.135; horizon main effect: F(1, 54)=1.839, p=.181, η^2=.033; Figure 5b).

When analysing the additional model parameter, we found that subjects had similar prior beliefs about bandits, given by the initial estimate of a bandit's mean (prior mean Q_0: F(2, 54)=.118, p=.889, η^2=.004; Figure 5c) and their uncertainty about it (prior variance σ_0: horizon main effect: F(1, 54)=.129, p=.721, η^2=.002; drug main effect: F(2, 54)=.06, p=.942, η^2=.002; drug-by-horizon interaction: F(2, 54)=2.162, p=.125, η^2=.074; WASI-by-horizon interaction: F(1, 54)=.022, p=.882, η^2<.001; Figure 5d). Interestingly, our dopamine manipulation seemed to affect this uncertainty in a gender-specific manner, with female subjects having larger values of σ_0 compared to males in the placebo group, and with the opposite being true in the amisulpride group (cf. Appendix 1). Taken together, these findings show that value-free random exploration was most sensitive to our drug manipulations.

Discussion

Solving the exploration-exploitation problem is non trivial, and one suggestion is that humans solve it using computationally demanding exploration strategies (1, 2), taking account of the uncertainty (variance) as well as the expected reward (mean) of each choice. Although tracking the distribution of summary statistics (e.g. mean and variance) is less resource costly than keeping track of full distributions (*52*), it nevertheless carries considerable costs when one has to keep track of multiple options, as in exploration. Indeed, in a three-bandit task such as that considered here, this results in a necessity to compute 6 key-statistics, drastically limiting computational resources when selecting among choice options (10). Real-life decisions often comprise an unlimited range of options, which results in a tracking of a multitude of key-statistics, potentially mandating a deployment of alternative more efficient strategies. Here, we demonstrate that two additional, less resource-hungry heuristics are at play during human decision-making, value-free random exploration and novelty exploration.

By assigning intrinsic value (novelty bonus (*42*)) to an option not encountered before (*53*), a novelty bonus can be seen as an efficient simplification of demanding algorithms, such as UCB (*3*, *4*). It is interesting to note that our winning model did not include UCB, but instead novelty exploration. This indicates humans might use such a novelty shortcut to explore unseen, or rarely visited, states to conserve computational costs when such a strategy is possible. A second exploration heuristic that also requires minimal computational resources, value-free random exploration, also plays a role in our task. Even though less optimal, its simplicity and neural plausibility renders it a viable strategy. We show through converging behavioural and modelling measures that both value-free random and novelty exploration were deployed in a goal-directed manner, coupled with increased levels of exploration when this was strategically useful.

Importantly, these heuristics were observed in all best models (1st, 2nd and 3rd position) even though each incorporated different exploration strategies. This suggests that the complex models made similar predictions in our task, and demonstrates that value-free random exploration is at play even when accounting for other value-based forms of random exploration (*1*, *7*), whether fixed or uncertainty-driven.

Exploration was captured in a similar manner to previous studies (*7*), by comparing in the same setting (i.e. same prior information) the first choice in a long decision horizon, where reward can be increased in the long term through information gain, and in a short decision horizon where information cannot subsequently be put to use. This means that by changing the opportunity to benefit from the information gained for the first sample, the long horizon invites extended exploration (7), what we find also in our study. This experimental manipulation is a well-established means for altering exploration and has been used extensively in previous studies (*7*, *21*, *34*, *54*). Nevertheless, there remains a possibility that a longer horizon may also affect the psychological nature of the task. In our task, reward outcomes were presented immediately after every draw, rendering it unlikely that perception of reward delays (i.e. delay discounting) is impacted. Moreover, a monetary bonus was given only at the end of the task, and thus did not impact a horizon manipulation. We also consider our manipulation was unlikely to change effort in each horizon, because the reward (i.e. size of the apple) remains the same at every draw, resulting in an equivalent reward-effort ratio (*55–58*). However, this issue can be addressed in further studies, for example, by equating the amount of button presses across both conditions.

Value-free random exploration might reflect other influences, such as attentional lapses or impulsive motor responses. We consider these as unlikely to a significant factor at play here. Indeed, there are two key features that would signify such effects. Firstly, these influences would be independent of task condition. Secondly, they would be expected to lead to shorter, or more variable, response latencies. In our data, we observe an increase in value-free exploration in the long horizon condition in both behavioural measures and model parameters, speaking against an explanation based upon simple mistakes. Moreover, we did not observe a difference in response latency for choices that were related to value-free random exploration (cf. Appendix 1), further arguing against mistakes. Lastly, the sensitivity of value-free random exploration to propranolol supports this being a separate process, and previous studies using the same drug did not find an effect on task mistakes (e.g. on accuracy (*59*); (*33*, *58–60*)). However, future studies could explore these exploration strategies in more detail including by reference to subjects' own self-reports.

It is still unclear how exploration strategies are implemented neurobiologically. Noradrenaline inputs, arising from the locus coeruleus (*63*) (LC) are thought to modulate exploration (*2*, *64*, *65*), though empirical data on its precise mechanisms and means of action remains limited. In this study, we found that noradrenaline impacted value-free random exploration, in contrast to novelty exploration and complex exploration. This might suggest that noradrenaline influences ongoing valuation or choice processes that discards prior information. Importantly, this effect was observed whether the complex exploration was an uncertainty-driven value-based random exploration (winning model), a directed exploration with value-based random exploration (2nd place model) or a combination of the above (3rd place model; cf. Appendix 1). This is consistent with findings in rodents where enhanced anterior cingulate noradrenaline release leads to more random behaviour (*32*). It is also consistent with pharmacological findings in monkeys that show enhanced choice consistency after reducing LC noradrenaline firing rates (*33*). It would be interesting for future studies to determine, in more detail, whether value-free random exploration is corrupting a value computation itself, or whether it exclusively biases the choice process.

We note that pupil diameter has been used as an indirect marker of noradrenaline activity (*66*), although the link between the two it not always straightforward (*36*). Because the effect of pharmacologically induced changes of noradrenaline levels on pupil size remains poorly understood (*36*, *67*), including the fact that previous studies found no effect of propranolol on pupil diameter (*36*, *68*), we opted against using pupillometry in this study. However, our current findings align with previous human studies that show an association between this indirect marker and exploration, but that study did not dissociate between the different potential exploration strategies that subjects could deploy (*69*). Future studies might usefully include indirect measures of noradrenaline activity, for example pupillometry, to examine a potential link between natural variations in noradrenaline levels and a propensity towards value-free random exploration.

The LC has two known modes of synaptic signalling (*63*), tonic and phasic, thought to have complementary roles (*31*). Phasic noradrenaline is thought to act as a reset button (*31*), rendering an agent agnostic to all previously accumulated information, a de facto signature of value-free random exploration. Tonic noradrenaline has been associated, although not consistently (*70*), with increased exploration (*64*, *71*), decision noise in rats (*72*) and more specifically with random as opposed to directed exploration strategies (*34*). This later study unexpectedly found that boosting noradrenaline decreased (rather than increased) random exploration, which the authors speculated was due to an interplay with phasic signalling. Importantly, the drug used in that study also affects dopamine function making it difficult to assign a precise interpretation to the finding. A consideration of this study influenced our decision to opt for drugs with high specificity for either dopamine or noradrenaline (*59*), enabling us to reveal highly specific effects on value-free random exploration. Although the contributions of tonic and phasic noradrenaline signalling cannot be disentangled in our study, our findings align with theoretical accounts and non-primate animal findings, indicating that phasic noradrenaline promotes value-free random exploration.

Aside from this 'reset signal' role, noradrenaline has been assigned other roles, including a role in memory function (*23*, *73*, *74*). To minimise a possible memory-related impact, we designed the task such that all necessary information was visible on the screen at all times. This means subjects did not have to memorise values for a given trial, rendering the task less susceptible to forgetting or other memory effects. Another role for noradrenaline relates to volatility and uncertainty estimation (*26–28*), as well as the energisation of behaviour (*24*, *25*). Non-human primates studies demonstrate a higher LC activation for high effort choices, suggesting that noradrenaline release facilitates energy mobilisation (*24*). Theoretical models also suggest that the LC is involved in the control of effort exertion. Thus, it is thought to contribute to trading off between effortful actions leading to large rewards and "effortless" actions leading to small rewards by modulating "raw" reward values as a function of the required effort (*25*). Our task can be interpreted as encapsulating such a trade-off: complex exploration strategies are effortful but optimal in terms of reward gain, while value-free random exploration requires little effort while occasionally leading to low reward. Applying this model, a noradrenaline boost could optimise cognitive effort allocation for high reward gain (*25*), thereby facilitating complex exploration strategies compared to value-free random exploration. In such a framework, blocking noradrenaline release should decrease usage of complex exploration strategies, leading to an increase of value-free random exploration which is the opposite of what we observed in our data. Another interpretation of an effort-facilitation model of noradrenaline is that a boost would help overcoming cost, i.e. the lack of immediate reward when selecting the low-value bandit, essentially providing a significant increase to the value of information gain. In line with our results, a decrease would interrupt this boost in valuation, removing an incentive to choose the low-value option. However, this theory is currently limited by the absence of empirical evidence for noradrenaline boosting valuation.

Noradrenaline blockade by propranolol has been shown previously to enhance metacognition (*75*), decrease information gathering (*59*), and attenuate arousal-induced boosts in incidental memory (*36*). All of these findings, including a decrease in value-free random exploration found here, suggests propranolol may influence how neural noise affects information processing. In particular, the results indicate that under propranolol behaviour is more deterministic and less influenced by ‘task-irrelevant’ distractions. This aligns with theoretical ideas, as well as recent optogenetic evidence (*32*), that propose noradrenaline infuses noise in a temporally targeted way (*31*). It also accords with studies implicating noradrenaline in attention shifts (for a review cf. (*76*)). Other theories of noradrenaline/catecholamine function can link to determinism (*64*, *65*), although the hypothesized direction of effect is different (i.e. noradrenaline increases determinism). This idea can be extended also to tasks where propranolol has been shown to attenuate a discrimination between different levels of loss (with no effect on the value-based exploration parameter, referred to in these studies as consistency) (*62*) and a reduction in loss aversion (*60*). This hints at additional roles for noradrenaline on prior information and task-distractibility during exploration in loss frame environments. Future studies investigating exploration in loss contexts might provide important additional information on these questions.

It is important to mention here that β-adrenergic receptors, the primary target of propranolol, have been shown (unlike α-adrenergic receptors) to increase synaptic inhibition within rat cortex (*77*), specifically through inhibitory GABA-mediated transmission (*78*).

Additionally β-adrenergic receptors are more concentrated in the intermediate layers in the prefrontal area (*79*), within which inhibition is favoured (*80*). Thus inhibitory mechanisms might account for noradrenaline-related task-distractibility and randomness, or the role of β-adrenergic receptors in executive function impairments (*81*). This raises the question of whether blocking β-adrenergic receptors might lead to an accumulation of synaptic noradrenaline, and therefore act via α-adrenergic receptors. To the best of our knowledge, evidence for such an effect is limited. A second question is whether the observed effects are a pure consequence of propranolol's impact on the brain, or whether they reflect peripheral effects of propranolol. When we examined peripheral markers (i.e. heart rate) and behaviour we found no evidence for an effect on any of our findings, rendering such influences unlikely. However, future studies using drugs that exclusively targets peripheral, but not central, noradrenaline receptors (e.g. (*82*)) are needed to answer this question conclusively.

Dopamine has been ascribed multiple functions besides reward learning (*83*), such as novelty seeking (*46*, *84*, *85*) or exploration in general (*43*). In fact, studies have demonstrated that there are different types of dopaminergic neurons in the ventral tegmental area, and that some contribute to non-reward signals, such as saliency and novelty (*44*). This suggests a role in novelty exploration. Moreover, dopamine has been suggested as important in an exploration-exploitation arbitration (*21*, *86*, *87*), although its precise role remains unclear, given reported effects on random exploration (*88*), on directed exploration (*45*, *89*), or no effects at all (*90*). A recent study found no effect following dopamine blockade using haloperidol (*87*), which interestingly also affects noradrenaline function (e.g. (*91*, *92*)). Our results did not demonstrate any main effect of dopamine manipulation on exploration strategies, even though blocking dopamine was associated with a trend level increase in exploitation (cf. Appendix 1). We believe it unlikely this reflects an ineffective drug dose as previous studies have found neurocognitive effects with the same dose (*36*, *59*, *93*, *94*).

One possible reason for an absence of significant findings is that our dopaminergic blockade targets D2/D3 receptors rather than D1 receptors, a limitation due a lack of available specific D1 receptor blockers for use in humans. An expectation of greater D1 involvement arises out of theoretical models (*95*) and a prefrontal hypothesis of exploration (*89*). Interestingly, we observed a weak gender-specific differential drug effect on subjects' uncertainty about an expected reward, with women being more uncertain than men in the placebo setting, but more certain in the dopamine blockade setting (cf. Appendix 1). This might be meaningful as other studies using the same drug have also found behavioural gender-specific drug effects (*96*). Upcoming, novel drugs (*97*) might be able help unravel a D1 contribution to different forms of exploration. Additionally, future studies could use approved D2/D3 agonists (e.g. ropinirole) in a similar design to probe further whether enhancing dopamine leads to a general increase in exploration.

In conclusion, humans supplement computationally expensive exploration strategies with less resource demanding exploration heuristics, and as shown here the latter include value-free random and novelty exploration. Our finding that noradrenaline specifically influences value-free random exploration demonstrates that distinct exploration strategies may be under specific neuromodulator influence. Our current findings may also be relevant to enabling a richer understanding of disorders of exploration, such as attention-deficit/hyperactivity disorder (*22*, *98*) including how aberrant catecholamine function might contribute to its core behavioural impairments.

Materials and Methods

Subjects

Sixty healthy volunteers aged 18 to 35 (mean =23.22, SD =3.615) participated in a double-blind, placebo-controlled, between-subjects study. The sample size was determined using power calculation taking effect sizes from our prior studies that used the same drug manipulations (*36*, *59*, *75*). Each subject was randomly allocated to one of three drug groups, controlling for an equal gender balance across all groups (cf. Appendix 1). Candidate subjects with a history of neurological or psychiatric disorders, current health issues, regular medications (except contraceptives), or prior allergic reactions to drugs were excluded from the study. Subjects had (self-reported) normal or corrected-to-normal vision. The groups consisted of 20 subjects each matched (cf. Appendix 2 Table 1) for gender and age. To evaluate peripheral drug effects, heart rate, systolic and diastolic blood pressure were collected to at three different time-points: 'at arrival', 'pre-task' and 'post-task', cf. Appendix 1 for details. At 50 minutes after administrating the 2nd drug, subjects were filled in the PANAS questionnaires (*50*) and completed the WASI Matrix Reasoning subtest (*49*). Subjects differed in mood (PANAS negative affect, cf. Appendix 1 for details) and marginally in intellectual abilities (WASI), and so we control for these potential confounders in our analyses (cf. Appendix 1 for uncorrected results). Subjects were reimbursed for their participation on an hourly basis and received a bonus according to their performance (proportional to the sum of all the collected apples' size). One subject from the amisulpride group was excluded due to not engaging in the task and performing at chance level. The study was approved by the UCL research ethics committee and all subjects provided written informed consent.

Pharmacological manipulation

To reduce noradrenaline functioning, we administered 40mg of the non-selective β-adrenoceptor antagonist propranolol 60 minutes before the task (Fig 1D). To reduce dopamine functioning, we administered 400mg of the selective D2/D3 antagonist amisulpride 90 minutes before the task. Because of different pharmacokinetic properties, drugs were administered at different times. Each drug group received the drug on its corresponding time point and a placebo at the other time point. The placebo group received placebo at both time points, in line with our previous studies (*36*, *59*, *75*).

Experimental paradigm

To quantify different exploration strategies, we developed a multi-armed bandit task implemented using Cogent (http://www.vislab.ucl.ac.uk/cogent.php) for MATLAB (R2018a). Subjects had to choose between bandits (i.e. trees) that produced samples (i.e. apples) with varying reward (i.e. size) in two different horizon conditions (Figure 1a-b). Bandits were displayed during the entire duration of a trial and there was no time limit for sampling from (choosing) the bandits. The sizes of apples they collected were summed and converted to an amount of juice (feedback), which was displayed during 2000 ms at the end of each trial. Subjects were instructed to endeavour to make the most juice and that they would receive a cash bonus proportional to their performance. Overall subjects received £10 per hour and a mean bonus of £1.12 (std: £0.06).

Similar to the horizon task (*7*), to induce different extents of exploration, we manipulated the horizon (i.e. number of apples to be picked: 1 in the short horizon, 6 in the long horizon) between trials. This horizon-manipulation, which has been extensively used to modulate exploratory behaviour (*21*, *34*, *54*, *99*), promotes exploration in the long horizon condition as there are more opportunities to gather reward.

Within a single trial, each bandit had a different mean reward μ (i.e. apple size) and associated uncertainty as captured by the number of initial samples (i.e. number of apples shown at the beginning of the trial). Each bandit (i.e. tree) i was from one of four generative processes (Figure 1c) characterised by different means μ_i and number of initial samples. The rewards (apple sizes) for each bandit were sampled from a normal distribution with mean μ_i, specific to the bandit, and with a fixed variance, S^2=0.8. The rewards were those sampled values rounded to the closest integer. Each distribution was truncated to [2, 10], meaning that rewards with values above or below this interval were excluded, resulting in a total of 9 possible rewards (i.e. 9 different apple sizes; cf. Figure 1 - Figure supplement 1 for a representation). The 'certain standard bandit' provided three initial samples and on every trial its mean μ_{cs} was sampled from a normal distribution: $\mu_{cs} \sim N(5.5, 1.4)$. The 'standard bandit' provided one initial sample and to make sure that its mean μ_s was comparable to μ_{cs}, the trials were split equally between the four following: $\{\mu_s = \mu_{cs} + 1;\ \mu_s = \mu_{cs} - 1;\ \mu_s = \mu_{cs} + 2;\ \mu_s = \mu_{cs} - 2\}$. The 'novel bandit' provided no initial samples and its mean μ_n was comparable to both μ_{cs} and μ_s by splitting the trials equally between the eight following: $\{\mu_n = \mu_{cs} + 1;\ \mu_n = \mu_{cs} - 1;\ \mu_n = \mu_{cs} + 2;\ \mu_n = \mu_{cs} - 2;\ \mu_n = \mu_s + 1;\ \mu_n = \mu_s - 1;\ \mu_n = \mu_s + 2;\ \mu_n = \mu_s - 2\}$. The 'low bandit' provided one initial sample which was smaller than all the other bandits' means on that trial: $\mu_l = min(\mu_{cs}, \mu_s, \mu_n) - 1$. We ensured that the initial sample from the low-value bandit was the smallest by resampling from each bandit in the trials were that was not the case. To make sure that our task captures heuristic exploration strategies, we simulated behaviour (cf. Figure 1). Additionally, in each trial, to avoid that some exploration strategies overshadow other ones, only three of the four different groups were available to choose from. Based on the mean of the initial samples, we identified the high-value option (i.e. the bandit with the highest expected reward) in trials where both the certain-standard and the standard bandit were present.

There were 25 trials of each of the four three-bandit combination making it a total of 100 different trials. They were then duplicated to measure choice consistency, defined as the frequency of making the same choice on identical trials (in contrast to a previous propranolol study where consistency was defined in terms of a value-based exploration parameter (*60*)). Each subject played these 200 trials both in a short and in a long horizon setting, resulting in a total of 400 trials. The trials were randomly assigned to one of four blocks and subjects were given a short break at the end of each of them. To prevent learning, the bandits' positions (left, middle or right) as well as their colour (8 sets of 3 different colours) where shuffled between trials. To ensure subjects distinguished different apple sizes and understood that apples from the same tree were always of similar size (generated following a normal distribution), they needed to undergo training prior to the main experiment. In training, based on three displayed apples of similar size, they were tasked to guess between two options, namely which apple was most likely to come from the same tree and then received feedback about their choice.

Statistical analyses

All statistical analyses were performed using the R Statistical Software (*100*). For computing ANOVA tests and pairwise comparisons the 'rstatix' package was used, and for computing effect sizes the 'lsr' package (*101*) was used. To ensure consistent performance across all subjects, we excluded one outlier subject (belonging to the amisulpride group) from our analysis due to not engaging in the task and performing at chance level (defined as randomly sampling from one out of three bandits, i.e. 33%). Each bandits' selection frequency for a horizon condition was computed over all 200 trials and not only over the trials where this specific bandit was present (i.e. 3/4 of 200 = 150 trials). In all the analysis comparing horizon conditions, except when looking at score values (Figure 2c), only the 1st draw of the long horizon was used. We compared behavioural measures and model parameters using (paired-samples) t-tests and repeated-measures (rm-) ANOVAs with a between-subject factor of drug group (propranolol group, amisulpride group, placebo group) and a within-subject factor horizon (long, short). Information seeking, expected values and scores were analysed using rm-ANOVAS with a within-subject factor horizon. Measures that were horizon-independent (e.g. prior mean), were analysed using one-way ANOVAs with a between-subject factor drug group. As drug groups differed in negative affect (cf. Appendix 2 Table 1), which, through its relationship to anxiety (*102*) is thought to affect cognition (*103*) and potentially exploration (*104*). We corrected for negative affect (PANAS) and IQ (WASI) in each analysis by adding those two measures as covariates in each ANOVA mentioned above (cf. Appendix 1 for analysis without covariates and analysis with physiological effect as an additional covariates). We report effect sizes using partial eta squared (η2) for ANOVAs and Cohen's d (d) for t-tests (*105*).

Computational modelling

We adapted a set of Bayesian generative models from previous studies (*1*), where each model assumed that different characteristics account for subjects' behaviour. The binary indicators $(\mathrm{c_{tr}}, \mathrm{c_n})$ indicate which components (value-free random and novelty exploration respectively) were included in the different models. The value of each bandit is represented as a distribution $N(Q, S)$ with $S = 0.8$, the sampling variance fixed to its generative value. Subjects have prior beliefs about bandits' values which we assume to be Gaussian with mean Q_0 and uncertainty σ_0. The subjects' initial estimate of a bandit's mean (Q_0; prior mean) and its uncertainty about it (σ_0; prior variance) are free parameters.

These beliefs are updated according to Bayes rule (detailed below) for each initial sample (note that there are no updates for the novel bandit).

Mean and variance update rules

At each time point t, in which a sample m, of one of the bandits is presented, the expected mean Q and precision $\tau = \frac{1}{\sigma^2}$ of the corresponding bandit i are updated as follows:

$$Q_{i,t+1} = \frac{\tau_{i,t} * \ Q_{i,t} + \tau_{samp} * m}{\tau_{i,t} + \ \tau_{samp}}$$

$$\tau_{t+1}^{i} = \tau_{samp} + \ \tau_{t}^{i}$$

where $\tau_{samp} = \frac{1}{S^2}$ is the sampling precision, with the sampling variance $S = 0.8$ fixed. Those update rules are equivalent to using a Kalman filter (*106*) in stationary bandits.

We examined three base models: the UCB model, the Thompson model and the hybrid model. The UCB model encompasses the UCB algorithm (captures directed exploration) and a softmax choice function (captures a value-based random exploration). The Thompson model reflects Thompson sampling (captures an uncertainty-driven value-based random exploration). The hybrid model captures the contribution of the UCB model and the Thompson model, essentially a mixture of the above. We computed three extensions of each model by either adding value-free random exploration $(\mathrm{c_{tr}}, \mathrm{c_n}) = (1{,}0)$, novelty exploration $(\mathrm{c_{tr}}, \mathrm{c_n}) = (0{,}1)$ or both heuristics $(\mathrm{c_{tr}}, \mathrm{c_n}) = (1{,}1)$, leading to a total of 12 models (see the labels on the x-axis in Figure 4a; $(\mathrm{c_{tr}}, \mathrm{c_n}) = (0{,}0)$ is the model with no extension). For additional models cf. Appendix 1. A coefficient c_{tr}=1 indicates that a ϵ-greedy component was added to the decision rule, ensuring that once in a while (every ϵ % of the time), another option than the predicted one is selected. A coefficient c_n=1 indicates that the novelty bonus η is added to the computation of the value of novel bandits and the Kronecker delta δ in front of this bonus ensures that it is only applied to the novel bandit. The models and their free parameters (summarised in Appendix 2 Table 5) are described in detail below.

Choice rules

UCB model. In this model, an information bonus γ is added to the expected reward of each option, scaling with the option's uncertainty (UCB). The value of each bandit i at timepoint t is:

$$V_{i,t} \ = Q_{i,t} + \gamma\sigma_{i,t} + c_n\eta\delta_{[i=novel]}$$

The probability of choosing bandit i was given by passing this into the softmax decision function:

$$P(c_t = i) = \frac{\mathrm{e}^{\beta V_{i,t}}}{\sum_x \mathrm{e}^{\beta V_{i,t}}} * (1 - c_{tr}\epsilon) + c_{tr}\frac{\epsilon}{3}$$

where β is the inverse temperature of the softmax (lower values producing more stochasticity), and the coefficient c_{tr} adds the value-free random exploration component.

Thompson model. In this model, based on Thompson sampling, the overall uncertainty can be seen as a more refined version of a decision temperature (*1*). The value of each bandit i is as before:

$$V_{i,t} = Q_{i,t} + c_n\eta\delta_{[\mathrm{i}=novel]}$$

A sample $x_{i,t} \sim N(V_{i,t}, \sigma_{i,t}^2)$ is taken from each bandit. The probability of choosing a bandit i depends on the probability that all pairwise differences between the sample from bandit i and the other bandits $j \neq i$ were greater or equal to 0 (see the probability of maximum utility choice rule (*107*)). In our task, because three bandits were present, two pairwise differences scores (contained in the two-dimensional vector u) were computed for each bandit. The probability of choosing bandit i is:

$$P(c_t = i) = P\big(\forall j: x_{i,t} > x_{j,t}\big) * (1 - c_{tr}\epsilon) + c_{tr}\frac{\epsilon}{3}$$

$$= \int_0^\infty \int_0^\infty \phi\big(u; M_{i,t}, C_{i,t}\big)\, du \; * (1 - c_{tr}\epsilon) + c_{tr}\frac{\epsilon}{3}$$

where ϕ is the multivariate Normal density function with mean vector

$$M_{i,t} = A_i \begin{pmatrix} V_{1,t} \\ V_{2,t} \\ V_{3,t} \end{pmatrix}$$

and covariance matrix

$$C_{i,t} = A_i \begin{pmatrix} \sigma_{1,t} & 0 & 0 \\ 0 & \sigma_{2,t} & 0 \\ 0 & 0 & \sigma_{3,t} \end{pmatrix} A_i^T$$

Where the matrix A_i computes the pairwise differences between bandit i and the other bandits. For example, for bandit $i = 1$:

$$A_1 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

Hybrid model. This model allows a combination of the UCB model and the Thompson model. The probability of choosing bandit i is:

$$P(c_t = i) = \Big(wP_{UCB}(c_t = i) + (1 - w)P_{Thompson}(c_t = i)\Big) * (1 - c_{tr}\epsilon) + c_{tr}\frac{\epsilon}{3}$$

where w specifies the contribution of each of the two models. P_{UCB} and $P_{Thompson}$ are calculated for c_{tr}=0. If w=1, only the UCB model is used while if w=0 only the Thompson model is used. In between values indicate a mixture of the two models.

All the parameters besides Q_0 and w were free to vary as a function of the horizon (cf. Appendix 2 Table 5) as they capture different exploration forms: directed exploration (information bonus γ; UCB model), novelty exploration (novelty bonus η), random exploration (inverse temperature β; UCB model), uncertainty-directed exploration (prior variance σ_0; Thompson model) and value-free random exploration (ϵ-greedy parameter). The prior mean Q_0 was fitted to both horizons together as we do not expect the belief of how good a bandit is to depend on the horizon. The same was done for w as assume the arbitration between the UCB model and the Thompson model does not depend on horizon.

Parameter estimation.

To fit the parameter values, we used the maximum a posteriori probability (MAP) estimate. The optimisation function used was fmincon in MATLAB. The parameters could vary within the following bounds: $\sigma_0 = [0.01, 6], Q_0 = [1, 10], \epsilon = [0, 0.5], \eta = [0, 5]$. The prior distribution used for the prior mean parameter Q_0 was the normal distribution: $Q_0 \sim N(5, 2)$ that approximates the generative distributions. For the ϵ-greedy parameter, the novelty bonus η and the prior variance parameter σ_0, a uniform distribution (of range equal to the specific parameters' bounds) was used, which is equivalent to performing MLE. A summary of the parameter values per group and per horizon can be found in Appendix 2 Table 6.

Model comparison.

We performed a K-fold cross-validation with $K = 10$. We partitioned the data of each subject (N_{trials} =400; 200 in each horizon) into K folds (i.e. subsamples). For model fitting in our model selection, we used maximum likelihood estimation (MLE), where we maximised the likelihood for each subject individually (fmincon was ran with 8 randomly chosen starting point to overcome potential local minima). We fitted the model using K-1 folds and validated the model on the remaining fold. We repeated this process K times, so that each of the K fold is used as a validation set once, and averaged the likelihood over held out trials. We did this for each model and each subject and averaged across subjects. The model with the highest likelihood of held-out data (the winning model) was the Thompson sampling with $(\mathrm{c_{tr}}, \mathrm{c_n}) = \{1,1\}$. It was also the model which accounted best for the largest number of subjects (Figure 4 – Figure supplement 1).

Parameter recovery.

To make sure that the parameters are interpretable, we performed a parameter recovery analysis. For each parameter, we took 4 values, equally spread, within a reasonable parameter range ($\sigma_0 = [0.5, 2.5], Q_0 = [1, 6], \epsilon = [0, 0.5], \eta = [0, 5]$). All parameters but Q_0 were free to vary as a function of the horizon. We simulated behaviour with one artificial agent for each 4^7 combinations using a new trial for each. The model was fitted using MAP estimation (cf. Parameter estimation) and analysed how well the generative parameters (generating parameters in Figure 5) correlated with the recovered ones (fitted parameters in Figure 5) using Pearson correlation (summarised in Figure 5c). In addition to the correlation we examined the spread (Figure 4 – Figure supplement 3) of the recovered parameters. Overall the parameters were well recoverable.

Model validation

To validate our model, we used each subjects' fitted parameters to simulate behaviour on our task (4000 trials per agent). The stimulated data (Figure 5 – Figure supplement 1), although not perfect, resembles the real data reasonably well. Additionally, to validate the behavioural indicators of the two different exploration heuristics we stimulated the behaviour of 200 agents using the winning model on one horizon condition (i.e. trials = 200). For the indicators of value-free random exploration, we stimulated behaviour with low ($\epsilon = 0$) and high ($\epsilon = 0.2$) values of the ϵ-greedy parameter. The other parameters were set to the mean parameter fits ($\sigma_0 = 1.312, \eta = 2.625, Q_0 = 3.2$). This confirms that higher amounts of value-free random exploration are captured by the proportion of low-value bandit selection (Figure 1f) and the choice consistency (Figure 1e). Similarly, for the indicator of novelty exploration, we simulated behaviour with low ($\eta = 0$) and high ($\eta = 2$) values of the novelty bonus η to validate the use of the proportion of the novel-bandit selection (Figure 1g). Again, the remaining parameters were set to the mean parameter fits ($\sigma_0 = 1.312, \epsilon = 0.1, Q_0 = 3.2$). Parameter values for high and low exploration were selected empirically from pilot and task data. Additionally, we simulated the effects of other exploration strategies in short and long horizon conditions (Figure 1 – Figure supplement 3-5). To simulate a long (versus short) horizon condition we increased the overall exploration by increasing other exploration strategies. Details about parameter values can be found in Appendix 2 Table 7.

Conflict of interest

The authors declare no competing financial interests.

Acknowledgements

M.D. is a predoctoral fellow of the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research. The participating institutions are the Max Planck Institute for Human Development and the University College London (UCL). T.U.H. is supported by a Wellcome Sir Henry Dale Fellowship (211155/Z/18/Z), a grant from the Jacobs Foundation (2017-1261-04), the Medical Research Foundation, a 2018 NARSAD Young Investigator Grant (27023) from the Brain and Behavior Research Foundation, and an ERC Starting Grant. R.J.D. holds a Wellcome Trust Investigator Award (098362/Z/12/Z). The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

References

1. S. J. Gershman, Deconstructing the human algorithms for exploration. *Cognition*. **173**, 34–42 (2018).
2. E. Schulz, S. J. Gershman, The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55**, 7–14 (2019).
3. P. Auer, Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2003).
4. A. Carpentier, A. Lazaric, M. Ghavamzadeh, R. Munos, P. Auer, Upper-confidence-bound algorithms for active learning in multi-armed bandits. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. **6925 LNAI**, 189–203 (2011).
5. P. Schwartenbeck, J. Passecker, T. U. Hauser, T. H. FitzGerald, M. Kronbichler, K. J. Friston, Computational mechanisms of curiosity and goal-directed exploration. *Elife* (2019), doi:10.7554/eLife.41703.
6. N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, R. J. Dolan, Cortical substrates for exploratory decisions in humans. *Nature*. **441**, 876–879 (2006).
7. R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, J. D. Cohen, Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074–2081 (2014).
8. W. R. Thompson, On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. **25**, 285–294 (1933).
9. J. D. Cohen, S. M. McClure, A. J. Yu, Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. B Biol. Sci.* **362**, 933–942 (2007).
10. I. C. Dezza, A. Cleeremans, W. Alexander, Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *J. Exp. Psychol. Gen.* (2019), doi:10.1037/xge0000546.
11. D. Papadopetraki, M. Froböse, A. Westbrook, B. Zandbelt, R. Cools, Quantifying the cost of cognitive stability and flexibility (2019), doi:10.1101/743120.
12. Z. Alexandre, S. Oleg, P. Giovanni, An information-theoretic perspective on the costs of cognition. *Neuropsychologia*. **123**, 5–18 (2019).
13. B. Wahn, P. König, Is attentional resource allocation across sensory modalities task-dependent? *Adv. Cogn. Psychol.* **13**, 83–96 (2017).
14. R. Marois, J. Ivanoff, Capacity limits of information processing in the brain. *Trends Cogn. Sci.* **9**, 296–305 (2005).
15. M. Botvinick, T. Braver, Motivation and cognitive control: From behavior to neural mechanism. *Annu. Rev. Psychol.* (2015), doi:10.1146/annurev-psych-010814-015044.
16. M. I. Froböse, A. Westbrook, M. Bloemendaal, E. Aarts, R. Cools, Catecholaminergic modulation of the cost of cognitive control in healthy older adults. *PLoS One*. **15**, 1–26 (2020).

17. W. Kool, J. T. McGuire, Z. B. Rosen, M. M. Botvinick, Decision Making and the Avoidance of Cognitive Demand. *J. Exp. Psychol. Gen.* **139**, 665–682 (2010).

18. R. Cools, The cost of dopamine for dynamic cognitive control. *Curr. Opin. Behav. Sci.* (2015), , doi:10.1016/j.cobeha.2015.05.007.

19. M. I. Froböse, R. Cools, Chemical neuromodulation of cognitive control avoidance. *Curr. Opin. Behav. Sci.* **22**, 121–127 (2018).

20. R. S. Sutton, A. G. Barto, Introduction to Reinforcement Learning. *MIT Press Cambridge* (1998), doi:10.1.1.32.7692.

21. W. K. Zajkowski, M. Kossut, R. C. Wilson, A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife*. **6**, 1–18 (2017).

22. T. U. Hauser, V. G. Fiore, M. Moutoussis, R. J. Dolan, Computational Psychiatry of ADHD: Neural Gain Impairments across Marrian Levels of Analysis. *Trends Neurosci.* **39**, 63–73 (2016).

23. S. J. Sara, A. Vankov, A. Hervé, Locus coeruleus-evoked responses in behaving rats: A clue to the role of noradrenaline in memory. *Brain Res. Bull.* **35**, 457–465 (1994).

24. C. Varazzani, A. San-Galli, S. Gilardeau, S. Bouret, Noradrenaline and dopamine neurons in the reward/effort trade-off: A direct electrophysiological comparison in behaving monkeys. *J. Neurosci.* **35**, 7866–7877 (2015).

25. M. Silvetti, E. Vassena, E. Abrahamse, T. Verguts, *Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner* (2018), vol. 14.

26. M. Silvetti, R. Seurinck, M. E. van Bochove, T. Verguts, The influence of the noradrenergic system on optimal control of neural plasticity. *Front. Behav. Neurosci.* **7**, 1–6 (2013).

27. A. J. Yu, P. Dayan, Uncertainty, neuromodulation, and attention. *Neuron*. **46**, 681–692 (2005).

28. M. R. Nassar, K. M. Rumsey, R. C. Wilson, K. Parikh, B. Heasly, J. I. Gold, Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat. Neurosci.* (2012), doi:10.1038/nn.3130.

29. J. David Johnson, Noradrenergic control of cognition: global attenuation and an interrupt function. *Med. Hypotheses* (2003), doi:10.1016/s0306-9877(03)00021-5.

30. S. Bouret, S. J. Sara, Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends Neurosci.* (2005), doi:10.1016/j.tins.2005.09.002.

31. P. Dayan, A. J. Yu, Phasic norepinephrine: A neural interrupt signal for unexpected events. *Netw. Comput. Neural Syst.* **17**, 335–350 (2006).

32. D. G. R. Tervo, M. Proskurin, M. Manakov, M. Kabra, A. Vollmer, K. Branson, A. Y. Karpova, Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*. **159**, 21–32 (2014).

33. C. I. Jahn, S. Gilardeau, C. Varazzani, B. Blain, J. Sallet, M. E. Walton, S. Bouret, Dual contributions of noradrenaline to behavioural flexibility and motivation, 2687–2702 (2018).

34. C. M. Warren, R. C. Wilson, N. J. Van Der Wee, E. J. Giltay, M. S. Van Noorden, J. D. Cohen, S. Nieuwenhuis, The effect of atomoxetine on random and directed exploration in humans. *PLoS One* (2017), doi:10.1371/journal.pone.0176034.

35. P. F. Fraundorfer, R. H. Fertel, D. D. Miller, D. R. Feller, Biochemical and pharmacological characterization of high-affinity trimetoquinol analogs on guinea pig and human beta adrenergic receptor subtypes: Evidence for partial agonism. *J. Pharmacol. Exp. Ther.* **270**, 665–674 (1994).

36. T. U. Hauser, E. Eldar, N. Purg, M. Moutoussis, R. J. Dolan, Distinct roles of dopamine and noradrenaline in incidental memory. *J. Neurosci.* (2019), doi:10.1523/jneurosci.0401-19.2019.

37. Ki Database, (available at https://pdsp.unc.edu/databases/pdsp.php).

38. N. Bunzeck, C. F. Doeller, R. J. Dolan, E. Duzel, Contextual interaction between novelty and reward processing within the mesolimbic system. *Hum. Brain Mapp.* (2012), doi:10.1002/hbm.21288.

39. B. C. Wittmann, N. D. Daw, B. Seymour, R. J. Dolan, Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron* (2008), doi:10.1016/j.neuron.2008.04.027.

40. S. J. Gershman, Y. Niv, Novelty and Inductive Generalization in Human Reinforcement Learning. *Top. Cogn. Sci.* **7**, 391–415 (2015).

41. H. Stojic, E. Shulz, P. P. Analytis, M. Speekenbrink, It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *PsyArXiv* (2018).

42. R. M. Krebs, B. H. Schott, H. Schütze, E. Düzel, The novelty exploration bonus and its attentional modulation. *Neuropsychologia* (2009), doi:10.1016/j.neuropsychologia.2009.01.015.

43. M. J. Frank, B. B. Doll, J. Oas-Terpstra, F. Moreno, Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat. Neurosci.* **12**, 1062–1068 (2009).

44. E. S. Bromberg-Martin, M. Matsumoto, O. Hikosaka, Dopamine in Motivational Control: Rewarding,

Aversive, and Alerting. *Neuron* (2010), , doi:10.1016/j.neuron.2010.11.022.

45. V. D. Costa, V. L. Tran, J. Turchi, B. B. Averbeck, Dopamine modulates novelty seeking behavior during decision making. *Behav. Neurosci.* (2014), doi:10.1037/a0037128.

46. E. Düzel, W. D. Penny, N. Burgess, Brain oscillations and memory. *Curr. Opin. Neurobiol.* (2010), , doi:10.1016/j.conb.2010.01.004.

47. K. Iigaya, T. U. Hauser, Z. Kurth-Nelson, J. P. O'Doherty, P. Dayan, R. J. Dolan, The value of what's to come: neural mechanisms coupling prediction error and reward anticipation. *bioRxiv* (2019), doi:10.1101/588699.

48. A. S. Kayser, J. M. Mitchell, D. Weinstein, M. J. Frank, Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology* (2015), doi:10.1038/npp.2014.193.

49. D. Wechsler, WASI -II: Wechsler abbreviated scale of intelligence - second edition. *J. Psychoeduc. Assess.* (2013), doi:10.1177/0734282912467756.

50. D. Watson, L. A. Clark, A. Tellegen, 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.

51. S. Agrawal, N. Goyal, Analysis of thompson sampling for the multi-armed bandit problem. *J. Mach. Learn. Res.* **23**, 1–26 (2012).

52. M. D'Acremont, P. Bossaerts, Neurobiological studies of risk assessment: A comparison of expected utility and mean-variance approaches. *Cogn. Affect. Behav. Neurosci.* **8**, 363–374 (2008).

53. N. C. Foley, D. C. Jangraw, C. Peck, J. Gottlieb, Novelty enhances visual salience independently of reward in the parietal lobe. *J. Neurosci.* (2014), doi:10.1523/JNEUROSCI.4171-13.2014.

54. C. M. Wu, E. Schulz, M. Speekenbrink, J. D. Nelson, B. Meder, Generalization guides human exploration in vast decision spaces. *Nat. Hum. Behav.* **2**, 915–924 (2018).

55. V. Skvortsova, S. Palminteri, M. Pessiglione, Learning to minimize efforts versus maximizing rewards: Computational principles and neural correlates. *J. Neurosci.* **34**, 15621–15630 (2014).

56. T. U. Hauser, E. Eldar, R. J. Dolan, Separate mesocortical and mesolimbic pathways encode effort and reward learning signals. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7395–E7404 (2017).

57. M. E. Walton, S. Bouret, What Is the Relationship between Dopamine and Effort? *Trends Neurosci.* **42**, 79–91 (2019).

58. J. D. Salamone, S. E. Yohn, L. López-Cruz, N. San Miguel, M. Correa, Activational and effort-related aspects of motivation: Neural mechanisms and implications for psychopathology. *Brain*. **139**, 1325–1347 (2016).

59. T. U. Hauser, M. Moutoussis, N. Purg, P. Dayan, R. J. Dolan, Beta-Blocker Propranolol Modulates Decision Urgency During Sequential Information Gathering. *J. Neurosci.* (2018), doi:10.1523/jneurosci.0192-18.2018.

60. P. Sokol-Hessner, S. F. Lackovic, R. H. Tobe, C. F. Camerer, B. L. Leventhal, E. A. Phelps, Determinants of Propranolol's Selective Effect on Loss Aversion. *Psychol. Sci.* **26**, 1123–1130 (2015).

61. D. Campbell-Meiklejohn, J. Wakeley, V. Herbert, J. Cook, P. Scollo, M. K. Ray, S. Selvaraj, R. E. Passingham, P. Cowen, R. D. Rogers, Serotonin and dopamine play complementary roles in gambling to recover losses. *Neuropsychopharmacology*. **36**, 402–410 (2011).

62. R. D. Rogers, M. Lancaster, J. Wakeley, Z. Bhagwagar, Effects of beta-adrenoceptor blockade on components of human decision-making. *Psychopharmacology (Berl).* **172**, 157–164 (2004).

63. J. Rajkowski, P. Kubiak, G. Aston-Jones, Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Res. Bull.* **35**, 607–616 (1994).

64. G. Aston-Jones, J. D. Cohen, AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).

65. D. Servan-Schreiber, H. Printz, J. D. Cohen, A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science (80-.).* (1990), doi:10.1126/science.2392679.

66. S. Joshi, Y. Li, R. M. Kalwani, J. I. Gold, Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron* (2016), doi:10.1016/j.neuron.2015.11.028.

67. S. Joshi, J. I. Gold, Pupil Size as a Window on Neural Substrates of Cognition. *Trends Cogn. Sci.* **24**, 466–480 (2020).

68. V. Koudas, A. Nikolaou, E. Hourdaki, S. G. Giakoumaki, P. Roussos, P. Bitsios, Comparison of ketanserin, buspirone and propranolol on arousal, pupil size and autonomic function in healthy volunteers. *Psychopharmacology (Berl).* (2009), doi:10.1007/s00213-009-1508-5.

69. M. Jepma, S. Nieuwenhuis, Pupil diameter predicts changes in the exploration-exploitation trade-off:

Evidence for the adaptive gain theory. *J. Cogn. Neurosci.* (2011), doi:10.1162/jocn.2010.21548.

70. Jepma, The role of the noradrenergic system in the exploration-exploitation trade-off: a pharmacological study. *Front. Hum. Neurosci.* (2010), doi:10.3389/fnhum.2010.00170.
71. M. Usher, J. D. Cohen, D. Servan-Schreiber, J. Rajkowski, G. Aston-Jones, The role of locus coeruleus in the regulation of cognitive performance. *Science (80-.).* **283**, 549–554 (1999).
72. G. A. Kane, E. M. Vazey, R. C. Wilson, A. Shenhav, N. D. Daw, G. Aston-Jones, J. D. Cohen, Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cogn. Affect. Behav. Neurosci.* **17**, 1073–1083 (2017).
73. Z. L. Rossetti, S. Carboni, Noradrenaline and dopamine elevations in the rat prefrontal cortex in spatial working memory. *J. Neurosci.* **25**, 2322–2329 (2005).
74. M. E. Gibbs, D. S. Hutchinson, R. J. Summers, Noradrenaline release in the locus coeruleus modulates memory formation and consolidation; roles for α- and β-adrenergic receptors. *Neuroscience*. **170**, 1209–1222 (2010).
75. T. U. Hauser, M. Allen, N. Purg, M. Moutoussis, G. Rees, R. J. Dolan, Noradrenaline blockade specifically enhances metacognitive performance. *Elife* (2017), doi:10.7554/eLife.24901.
76. I. Trofimova, T. W. Robbins, Temperament and arousal systems: A new synthesis of differential psychology and functional neurochemistry. *Neurosci. Biobehav. Rev.* **64**, 382–402 (2016).
77. B. D. Waterhouse, H. C. Moises, H. H. Yeh, D. J. Woodward, Norepinephrine enhancement of inhibitory synaptic mechanisms in cerebellum and cerebral cortex: Mediation by beta adrenergic receptors. *J. Pharmacol. Exp. Ther.* **221**, 495–506 (1982).
78. B. D. Waterhouse, H. C. Moises, H. H. Yeh, H. M. Geller, D. J. Woodward, Comparison of norepinephrine- and benzodiazepine-induced augmentation of Purkinje cell response to γ-aminobutyric acid (GABA). *J. Pharmacol. Exp. Ther.* **228**, 257–267 (1984).
79. P. S. Goldman-Rakic, M. S. Lidow, D. W. Gallager, Overlap of dopaminergic, adrenergic, and serotoninergic receptors and complementarity of their subtypes in primate prefrontal cortex. *J. Neurosci.* **10**, 2125–2138 (1990).
80. J. S. Isaacson, M. Scanziani, How Inhibition Shapes Cortical Activity Excitation and inhibition walk hand in hand. *Neuron*. **72**, 231–243 (2011).
81. H. Salgado, M. Treviño, M. Atzori, Layer- and area-specific actions of norepinephrine on cortical synaptic transmission. *Brain Res.* **1641**, 163–176 (2016).
82. B. De Martino, B. A. Strange, R. J. Dolan, Noradrenergic neuromodulation of human attention for emotional and neutral stimuli. *Psychopharmacology (Berl).* **197**, 127–136 (2008).
83. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science (80-.).* **275**, 1593–1599 (1997).
84. B. C. Wittmann, N. D. Daw, B. Seymour, R. J. Dolan, Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron*. **58**, 967–973 (2008).
85. V. D. Costa, V. L. Tran, J. Turchi, B. B. Averbeck, Dopamine modulates novelty seeking behavior during decision making. *Behav. Neurosci.* **128**, 556–566 (2014).
86. A. S. Kayser, J. M. Mitchell, D. Weinstein, M. J. Frank, Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology*. **40**, 454–462 (2015).
87. K. Chakroun, D. Mathar, A. Wiehler, F. Ganzer, J. Peters, Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *bioRxiv*, 706176 (2019).
88. F. Cinotti, V. Fresno, N. Aklil, E. Coutureau, B. Girard, A. R. Marchand, M. Khamassi, Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Sci. Rep.* **9**, 1–14 (2019).
89. M. J. Frank, B. B. Doll, J. Oas-Terpstra, F. Moreno, Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat. Neurosci.* (2009), doi:10.1038/nn.2342.
90. L. K. Krugel, G. Biele, P. N. C. Mohr, S. C. Li, H. R. Heekeren, Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proc. Natl. Acad. Sci. U. S. A.* (2009), doi:10.1073/pnas.0905191106.
91. J. Fang, P. H. Yu, Effect of haloperidol and its metabolites on dopamine and noradrenaline uptake in rat brain slices. *Psychopharmacology (Berl).* (1995), doi:10.1007/BF02246078.
92. M. Toru, M. Takashima, Haloperidol in large doses reduces the cataleptic response and increases noradrenaline metabolism in the brain of the rat. *Neuropharmacology* (1985), doi:10.1016/0028-3908(85)90079-6.
93. T. Kahnt, S. C. Weber, H. Haker, T. W. Robbins, P. N. Tobler, Dopamine D2-Receptor Blockade Enhances Decoding of Prefrontal Signals in Humans. *J. Neurosci.* (2015), doi:10.1523/jneurosci.4182-14.2015.

94. T. Kahnt, P. N. Tobler, Dopamine Modulates the Functional Organization of the Orbitofrontal Cortex. *J. Neurosci.* (2017), doi:10.1523/jneurosci.2827-16.2016.

95. M. D. Humphries, M. Khamassi, K. Gurney, Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci.* (2012), doi:10.3389/fnins.2012.00009.

96. A. Soutschek, C. J. Burke, A. Raja Beharelle, R. Schreiber, S. C. Weber, I. I. Karipidis, J. Ten Velden, B. Weber, H. Haker, T. Kalenscher, P. N. Tobler, The dopaminergic reward system underpins gender differences in social preferences. *Nat. Hum. Behav.* (2017), doi:10.1038/s41562-017-0226-y.

97. A. Soutschek, G. Gvozdanovic, R. Kozak, S. Duvvuri, N. de Martinis, B. Harel, D. L. Gray, E. Fehr, A. Jetter, P. N. Tobler, Dopaminergic D1 Receptor Stimulation Affects Effort and Risk Preferences. *Biol. Psychiatry* (2019), doi:10.1016/j.biopsych.2019.09.002.

98. T. U. Hauser, R. Iannaccone, J. Ball, C. Mathys, D. Brandeis, S. Walitza, S. Brem, Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry*. **71**, 1165–1173 (2014).

99. D. Guo, A. J. Yu, in *Advances in Neural Information Processing Systems* (2018).

100. R. R Development Core Team, *R: A Language and Environment for Statistical Computing* (2011).

101. D. Navarro, *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5)* (2015; http://ua.edu.au/ccs/teaching/lsr).

102. D. Watson, L. A. Clark, G. Carey, Positive and Negative Affectivity and Their Relation to Anxiety and Depressive Disorders. *J. Abnorm. Psychol.* **97**, 346–353 (1988).

103. S. J. Bishop, C. Gagne, Anxiety, Depression, and Decision Making: A Computational Perspective. *Annu. Rev. Neurosci.* **41**, 371–388 (2018).

104. L. de Visser, L. J. van der Knaap, A. J. A. E. van de Loo, C. M. M. van der Weerd, F. Ohl, R. van den Bos, Trait anxiety affects decision-making differently in healthy men and women: Towards gender-specific endophenotypes of anxiety. *Neuropsychologia*. **48**, 1598–1606 (2010).

105. J. T. E. Richardson, Eta squared and partial eta squared as measures of effect size in educational research. *Educ. Res. Rev.* (2011), , doi:10.1016/j.edurev.2010.12.001.

106. C. M. Bishop, in *Information Science and Statistics* (2006).

107. M. Speekenbrink, E. Konstantinidis, Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* (2015), doi:10.1111/tops.12145.

Appendix 1

Drug effect on response times

There were no differences in response times (RT) between drug groups in the one-way ANOVA. Neither in the mean RT (ANOVA: F(2, 54)=1.625, p=.206, η2=.057) nor in its variability (standard deviation; F(2, 54)=1.85, p=.16, η2=.064).

Bandit effect on response times

There was no difference in response times between bandits in the repeated-measures ANOVA (bandit main effect: F(1.78 , 99.44)=1.634 , p=.203 , η2=.028; Figure 3 – Figure supplement 1).

Horizon effect on response times

There were no differences in RT between horizon conditions in the repeated-measures ANOVA with the between-subject factor drug group, the within-subject factor horizon condition and the covariates WASI and PANAS negative score (horizon main effect: F(1, 54)=1.443, p=.235, η2=.026; drug main effect: F(2 , 54)=1.625, p=.206, η2=.057; drug-by-horizon interaction: F(2, 54)=.431, p=.652, η2=.016. In the long horizon, the RT decreased with each sample (sample main effect: F(1.36, 73.5)=13.626, p<.001, η2=0.201; Pairwise comparisons: sample 1 vs 2: t(59)=20.968, p<.001, d=2.73; sample 2 vs 3: t(59)=11.825, p<.001, d=1.539; sample 3 vs 4: t(59)=7.862, p<.001, d=1.024; sample 4 vs 5: t(59)=4.117, p<.001, d=1.539; sample 5 vs 6: t(59)=2.646, p=.01, d=1.024; Figure 2 – Figure supplement 1b).

PANAS

The Positive Affect and Negative Affect scale (PANAS; (*50*)) was completed 50 minutes after the 2nd drug administration and 10 minutes prior to the task. Groups had similar positive affect but differed in negative affect (cf. Appendix 2 Table 1), driven by a higher score in the placebo group (pairwise comparisons: placebo vs propranolol: t(56)=2.801, p=.007, d=.799; amisulpride vs placebo: t(56)=-2.096, p=.041, d=.557; amisulpride vs propranolol: t(56)=.669, p=.506, d=.383). It is unclear whether this difference was driven by the drug manipulation, but similar studies have not reported such an effect (e.g. (*36*, *59*, *61*, *62*, *75*)). We controlled for a possible influence of these measures in all our analyses.

Physiological effects

Heart rate, systolic and diastolic pressure were obtained at 3 time points: at the beginning of the experiment before giving the drug ('at arrival'), after giving the drug just before the task ('pre-task'), and after finishing task and questionnaires ('post-task'). The post-task heart rate was lower for participants who received propranolol compared to the other 2 groups (1-way ANOVA: F(2, 55)=7.249, p=.002, η^2=.209; cf. Appendix 2 Table 2). A two-way ANOVA with the between-subject factor of drug group and within-subject factor of time (all three time points), showed a time-dependent decrease in heart rate (F(1.74, 95.97)=99.341, p<.001, η^2= .644), in systolic pressure (F(2, 110)=8.967, p<.001, η^2=.14) and in diastolic pressure (F(2, 110)=.874, p=.42, η^2=.016), indicating subjects relaxed across the course of the study. Those reductions did not differ between drug group (drug main effect: heart rate: F(2, 55)=1.84, p=.169, η^2=.063; systolic pressure: F(2, 55)=1.08, p=.347, η^2=.038; diastolic pressure: F(2, 55)=.239, p=.788, η^2=.009; drug-by-time interaction: heart rate: F(3.49, 95.97)=1.928, p=.121, η^2=.066; systolic pressure: F(4, 110)=1.6, p=.179, η^2=.055; diastolic pressure: F(4, 110)=.951, p=.438, η^2=.033).

Task performance score

The performance did not differ between drug groups (total score: drug main effect: F(2 , 5)=2.313, p=.109, η2=.079) but it was increased in subjects with higher IQ scores (WASI main effect: F(1 , 54)=17.172, p<.001, η2=.241).

In the long horizon, the score increased with each sample (sample main effect: F(3.12, 174.97)=103.469, p<.001, η2=0.649; Pairwise comparisons: sample 1 vs 2: t(59)=-6.737, p<.001, d=0.877; sample 2 vs 3: t(59)=-3.69, p<.001, d=0.48; sample 3 vs 4: t(59)=-5.167, p<.001, d=0.673; sample 4 vs 5: t(59)=-2.832, p=.006, d=0.48; sample 5 vs 6: t(59)=-2.344, p=.022, d=0.673; Figure 2 – Figure supplement 1a). The increase in reward was larger in trials where the first draw was exploratory (linear regression slope coefficient: mean=0.118, sd=0.038) compared to when it was exploitative (linear regression slope coefficient: mean=0.028, sd=0.041; t-tests for slope coefficients: t(58)=-12.161, p<.001, d=-1.583; Figure 2 - Figure supplement 1d), suggesting that exploration was used beneficially and subjects benefitted from their initial exploration.

Dopamine effect on high-value bandit sampling frequency

The amisulpride group had a marginal tendency towards selecting the high-value bandit, meaning that they were disposed to exploit more overall (propranolol group excluded: horizon main effect: F(1, 35)=3.035, p=.09, η2=.08; drug main effect: F(1, 35)=3.602, p=.066, η2=.093; drug-by-horizon interaction: F(1, 35)=2.15, p=.151, η2=.058). This trend effect was not observed when all 3 groups were included (horizon main effect: F(1, 54)=3.909, p=.053, η2=.068; drug main effect: F(2, 54)=1.388, p=.258, η2=.049; drug-by-horizon interaction: F(2, 54)=.834, p=.44, η2=.03).

Gender effects

When adding gender as a between-subjects variable in the repeated-measures ANOVAs, none of the main results changed. Interestingly, we observed a drug-by-gender interaction in the prior variance σ_0 (drug-by-gender interaction: F(2, 51)=5.914, p=.005, η2=.188; Figure 5 – Figure supplement 2), driven by the fact that, female subjects in the placebo group had a larger average σ_0 (across both horizon conditions) compared to males (t(20)=2.836, p=.011, d=1.268), whereas male subjects have a larger σ_0 compared to females in the amisulpride group, (t(19)=-2.466, p=.025, d=1.124; propranolol group: t(20)=-0.04, p=.969, d=.018). This suggests that in a placebo setting, females are on average more uncertain about an option's expected value, whereas in a dopamine blockade setting males are more uncertain. Besides this effect, we observed a trend-level significance in response times (RT), driven primarily by female subjects tending to have a faster RT in the long horizon compared to male subjects (gender main effect: F(1, 51)=3.54, p=.066, η2=.065).

Horizon and drug effects without covariate

When analysing the results without correcting for IQ (WASI) and negative affect (PANAS), similar results are obtained. The high-value bandit is picked more in the short-horizon condition indicating exploitation (F(1, 56)=44.844, p<.001, η2=.445), whereas the opposite phenomenon is observed in the low-value bandit (F(1, 56)=24.24, p<.001, η2=.302) and the novel bandit (horizon main effect: F(1, 56)=30.867, p<.001, η2=.355), indicating exploration. In line with these results, the model parameters for value-free random exploration (ϵ: F(1, 56)=10.362, p=.002, η2=.156) and novelty exploration (η: F(1, 56)=38.103, p<.001, η2=.405) are larger in the long compared to the short horizon condition. Additionally, noradrenaline blockade reduces value-free random exploration as can be seen in the two behavioural signatures, frequency of picking the low-value bandit (F(2, 56)=2.523, p=.089, η2=.083; Pairwise comparisons: placebo vs propranolol: t(40)=2.923, p=.005, d=.654; amisulpride vs placebo: t(38)=-.587, p=.559, d=.133; amisulpride vs propranolol: t(38)=2.171, p=.034, d=.496), and in the consistency (F(2, 56)=3.596, p=.034, η2=.114; Pairwise comparisons: placebo vs propranolol: t(40)=-3.525, p=.001, d=.788; amisulpride vs placebo: t(38)=1.107, p=.272, d=.251; amisulpride vs propranolol: t(38)=-2.267, p=.026, d=.514), as well as in the model parameter for value-free random exploration (ϵ: F(2, 56)=3.205, p=.048, η2=.103; Pairwise comparisons: placebo vs propranolol: t(40)=3.177, p=.002, d=.71; amisulpride vs placebo: t(38)=.251, p=.802, d=.057; amisulpride vs propranolol: t(38)=2.723, p=.009, d=.626).

Horizon and drug effects with heart rate as covariate

When analysing results but now correcting for the post-experiment heart rate (cf. Appendix 2 Table 1) in addition to IQ (WASI) and negative affect (PANAS), we obtained similar results. Noradrenaline blockade reduced value-free random exploration as seen in two behavioural signatures, frequency of picking the low-value bandit (F(2, 52)= 4.014, p=.024, η^2=.134; Pairwise comparisons:(placebo vs propranolol: t(40)= 2.923, p=.005, d=.654; amisulpride vs propranolol: t(38)= 2.171, p=.034, d=.496; amisulpride vs placebo: t(38)= -.587, p=.559, d=.133), and consistency (F(2, 52)= 5.474, p=.007, η^2=.174; Pairwise comparisons: placebo vs propranolol: t(40)= -3.525, p=.001, d=.788; amisulpride vs propranolol: t(38)= -2.267, p=.026, d=.514; amisulpride vs placebo: t(38)= 1.107, p=.272, d=.251), as well as in a model parameter for value-free random exploration (ϵ: F(2, 52)= 4.493, p=.016, η^2=.147; Pairwise comparisons: placebo vs propranolol: t(40)= 3.177, p=.002, d=.71; amisulpride vs propranolol: t(38)= 2.723, p=.009, d=.626; amisulpride vs placebo: t(38)=.251, p=.802, d=.057).

Other model results

When analysing the fitted parameter values of both the 2nd winning model (UCB + ϵ + η) and 3rd winning model (hybrid + ϵ + η), similar results pertain. Thus, a value-free random exploration parameter was reduced following noradrenaline blockade in the 2nd winning model (ϵ: F(2, 54)=4.503, p=.016, η^2=.143; Pairwise comparisons: placebo vs propranolol: t(38)=2.185, p=.033, d=.386; amisulpride vs propranolol: t(40)=1.724, p=.089, d=.501; amisulpride vs placebo: t(40)=-.665, p=.508, d=.151) and was affected at a trend-level significance in the 3rd winning model (ϵ: F(2, 54)=3.04, p=.056, η^2=.101). These results highlight our finding that value-free random exploration is modulated by noradrenaline and additionally demonstrates this is independent of the complex exploration strategy used as well as the value function.

Bandit combination effect

Behavioural results were analysed additionally for each bandit combination separately. The high-value bandit was picked more when there was no novel bandit (pairwise comparisons: [certain-standard, standard, low] vs [certain-standard, standard, novel]: t(59)=-15.122, p<.001, d=1.969 ; [certain-standard, standard, low] vs [certain-standard, novel, low]: t(59)=12.905, p<.001, d=1.68; [certain-standard, standard, low] vs [standard, novel, low]: t(59)=18.348, p<.001, d=2.389), and less when its value was less certain ([standard, novel, low] vs [certain-standard, standard, novel]: t(59)=6.986, p<.001, d=.909; [standard, novel, low] vs [certain-standard, novel, low] : t(59)=5.44, p<.001, d=.708; bandit combination main effect: F(1.81, 101.33)=237.051, p<.001, η^2=.809; [certain-standard, standard, novel] vs [certain-standard, novel, low]: t(59)=.364, p=.717, d=.047; Figure 3 – Figure supplement 2a). The novel bandit was picked the most when the high-value bandit was less certain, then when the high-value bandit was more certain and it was picked the least when both certain and certain standard bandits were present ([standard, novel, low] vs [certain-standard, novel, low]: t(59)=-5.001, p<.001, d=.651; [standard, novel, low] vs [certain-standard, standard, novel]: t(59)=-9.414, p<.001, d=1.226; [certain-standard, novel, low] vs [certain-standard, standard, novel]: t(59)=-4.146, p<.001, d=.54; bandit combination main effect: F(2, 112)=42.44, p<.001, η^2=.431; Figure 3 – Figure supplement 2b). The low-value bandit was picked less when the high-value bandit was more certain ([certain-standard, novel, low] vs [certain-standard, standard, low]: t(59)=2.731, p=.008, d=.356; [certain-standard, novel, low] vs [standard, novel, low]: t(59)=-1.958, p=.055, d=.255; bandit combination main effect: F(1.66, 92.74)=4.534, p=.019, η2=.075; [certain-standard, standard, low] vs [standard, novel, low]: t(59)=1.32, p=.192, d=.172; Figure 3 – Figure supplement 2c).

Other effects on choice consistency

Our results demonstrate a drug-by-horizon interaction on choice consistency (F(2, 54)=3.352, p=.042, η^2=.110; Figure 3c), mainly driven by the fact that frequency of selecting the same option is increased in the long (compared to the short) horizon in the amisulpride group, while there is no significant horizon difference in the other two drug groups (pairwise comparison for horizon effect: amisulpride group: t(19)=2.482, p=.023, d=.569; propranolol group: t(20)=-1.91, p=.071, d=.427; placebo group: t(20)=.505, p=.619, d=.113). It is not entirely clear why catecholamines would increase the differentiation between the horizon conditions and this relatively weak effect should be replicated before interpreting.

Stand-alone heuristic models

We also analysed stand-alone heuristic models, in which there is no value computation (value of each bandit i: $V_i = 0$). The held-out data likelihood for such heuristic model combined with novelty exploration had a mean of m=0.367 (sd=0.005). The model in which we added value-free random exploration on top of novelty exploration had a mean of m=0.384 (sd=0.006). These models performed poorly, although better than chance level. Importantly, adding value-free random exploration improved performance. This highlights that subjects' combine complex and heuristic modules in exploration.

Appendix 2

	Propranolol	Placebo	Amisulpride	
Gender (M/F)	10/10	10/10	10/9	
Age	22.80 (3.59)	23.80 (4.23)	23.05 (3.01)	F(2,56)=.404, p=.669, η^2=.014
Intellectual abilities	22.8 (1.85)	22.6 (3.70)	24.37 (2.45)	F(2,56)=2.337, p=.106, η^2=.077
Positive affect	24.55 (8.99)	28.90 (7.56)	29.58 (10.21)	F(2,56)=1.832, p=.170, η^2=.061
Negative affect	10.65 (.81)	12.75 (3.63)	11.16 (1.71)	F(2,56)=4.259, p=.019, η^2=.132

Appendix 2 Table 1.

Characteristics of drug groups. The drug groups did not differ in gender, age, nor in intellectual abilities (adapted WASI matrix test). Groups differed in negative affect (PANAS), driven by a higher score in the placebo group (pairwise comparisons: placebo vs propranolol: t(56)=2.801, p=.007, d=.799; amisulpride vs placebo: t(56)=-2.096, p=.041, d=.557; amisulpride vs propranolol: t(56)=.669, p=.506, d=.383). For more details cf. Appendix 1. Mean (SD).

		Propranolol	Placebo	Amisulpride	
Heart rate (BPM)	At arrival	74.9 (10.8)	77,2 (12,6)	77.7 (13.8)	F(2, 55)=.290, p=.749, η^2=.010
	Pre-task	62,6 (8,5)	65,8 (8,3)	64,6 (9,8)	F(2, 55)=.667, p=.517, η^2=.024
	Post-task	55,7 (6,7)	64,4 (6,9)	63,4 (10,0)	F(2, 55)=7.249, p=.002, η^2=.209
Systolic blood pressure	At arrival	117,2 (10,4)	115,0 (9,7)	117,9 (9,7)	F(2, 55)=.438, p=.648, η^2=.016
	Pre-task	109,4 (9,2)	111,8 (8,6)	114,9 (8,6)	F(2, 55)=1.841, p=.168, η^2=.063
	Post-task	109,5 (8,2)	113,9 (11,3)	114,6 (9,3)	F(2, 55)=1.584, p=.214, η^2= .054
Diastolic blood pressure	At arrival	71,5 (7,8)	71,2 (6,7)	72,3 (6,7)	F(2, 55)=.115, p=.891, η^2=.004
	Pre-task	68,3 (7,0)	71,1 (10,6)	72,0 (5,9)	F(2, 55)=1.111, p=.337, η^2= .039
	Post-task	70,8 (7,3)	70,9 (8,0)	70,3 (6,6)	F(2, 55)=.037, p=.964, η^2=.001

Appendix 2 Table 2.

Physiological effects on drug groups. The drug groups also differed in post-experiment heart rate, driven by lower values in the propranolol group (pairwise comparisons: placebo vs propranolol: t(55)=3.5, p=.001, d=1.293; amisulpride vs placebo: t(55)= -.394, p=.695, d=.119 ; amisulpride vs propranolol: t(55)=3.013, p=.004, d=.921). For detailed statistics and analysis accounting for this cf. Appendix 1. Mean (SD).

	Horizon	Mean (sd)	Two-way repeated-measures ANOVA
			Main effect of horizon
Expected value	**short**	6.368 (0.335)	F(1, 56)=19.457, p<.001, η^2=.258
	long	6.221 (0.379)	
Initial samples	**short**	1.282 (0.247)	F(1, 56)=58.78, p<.001, η^2=.512
	long	1.084 (0.329)	
Score (1st sample)	**short**	5.904 (0.192)	F(1, 56)=58.78, p<.001, η^2=.512
	long	5.82 (0.182)	
Score (average)	**short**	5.904 (0.192)	F(1, 56)=103.759, p<.001, η^2=.649
	long	6.098 (0.222)	

Appendix 2 Table 3.

Table of statistics and behavioural values of Figure 2. All of those measures were modulated by the horizon condition.

		Mean (sd)			**Two-way repeated-measures ANOVA**			
	Horizon	Amisulpride	Placebo	Propranolol	**Main effect**		**Interaction**	
High-value bandit	**short**	54.55 (8.87)	49.38 (9.10)	50.98 (11.4)	**D**	F(2, 54)=1.388, p=.258, η^2=.049	**DH**	F(2, 54)=.834, p=.440, η^2=.030
	long	41.90 (8.47)	44.10 (13.88)	41.90 (13.57)	**H**	F(1, 54)=3.909, p=.053, η^2=.068	**HW**	F(1, 54)=13.304, p=.001, η^2=.198
Low-value bandit	**short**	3.32 (2.33)	4.28 (2.98)	2.50 (2.48)	**D**	F(2, 54)=7.003, p=.002, η^2=.206	**DH**	F(2, 54)=2.154, p=.126, η^2=.074
	long	5.45 (3.76)	5.35 (3.40)	3.45 (2.18)	**H**	F(1, 54)=4.069, p=.049, η^2=.070	**HW**	F(1, 54)=1.199, p=.278, η^2=.022
Novel bandit	**short**	36.87 (9.49)	39.02 (10.94)	40.15 (12.43)	**D**	F(2, 54)=1.498, p=.233, η^2=.053	**DH**	F(2, 54)=.542, p=.584, η^2=.020
	long	46.82 (12.1)	43.62 (16.27)	48.55 (16.59)	**H**	F(1, 54)=5.593, p=.022, η^2=.094	**HW**	F(1, 54)=13.897, p<.001, η^2=.205
Consistency	**short**	64.16 (12.27)	62.70 (12.59)	73.00 (11.33)	**D**	F(2, 54)=7.154, p=.002, η^2=.209	**DH**	F(2, 54)=3.352, p=.042, η^2=.110
	long	68.11 (10.34)	64.00 (8.93)	70.55 (9.91)	**H**	F(1, 54)=1.333, p=.253, η^2=.024	**HW**	F(1, 54)=.409, p=.525, η^2=.008

Appendix 2 Table 4.

Table of statistics and behavioural measure values of Figure 3. The drug groups differed in low-value bandit picking frequency (pairwise comparisons: placebo vs propranolol: t(40)=2.923, p=.005, d=.654; amisulpride vs placebo: t(38)=-.587, p=.559, d=.133; amisulpride vs propranolol: t(38)=2.171, p=.034, d=.496) and choice consistency (placebo vs propranolol: t(40)=-3.525, p=.01, d=.788; amisulpride vs placebo: t(38)=1.107, p=.272, d=.251; amisulpride vs propranolol: t(38)=-2.267, p=.026, d=.514). The main effect is either of drug group (D) or of horizon (H). The interaction is either drug-by-horizon (DH) or horizon-by-WASI (measure of IQ; HW).

	Model	**Thompson**				**UCB**				**Hybrid**			
			$+\epsilon$	$+\eta$	$+\epsilon$ $+\eta$		$+\epsilon$	$+\eta$	$+\epsilon$ $+\eta$		$+\epsilon$	$+\eta$	$+\epsilon+\eta$
Parameters	Horizon independent	Q_0	Q_0	Q_0	Q_0	Q_0	Q_0	Q_0	Q_0	w, Q_0	w, Q_0	w, Q_0	w, Q_0
	Horizon dependent	σ_0	σ_0, ϵ	σ_0, η	$\sigma_0, \epsilon,$ η	γ, β	$\gamma, \beta,$ ϵ	$\gamma, \beta,$ η	$\gamma, \beta,$ ϵ, η	$\sigma_0, \gamma,$ β	$\sigma_0, \gamma,$ β, ϵ	$\sigma_0, \gamma,$ β, η	$\sigma_0, \gamma,$ $\beta,$ ϵ, η
Model selection	Mean held-out data likelihood	50.2 (8.1)	52.7 (7.1)	52,2 (8.7)	55.3 (8.4)	52.9 (8.0)	52.9 (8.0)	53.4 (8.1)	55.1 (8.8)	53.5 (8.1)	53.8 (8.4)	55.0 (8.4)	55 5.1 (8.5)
	Subjects' for which model fits best (out of 12)	0	3	2	20	0	0	1	20	0	0	7	6
	Subjects' for which model fits best (out of 3 best)	-	-	-	27	-	-	-	22	-	-	-	10

Appendix 2 Table 5.

Table of parameters used for each model compared during model selection (Figure 4). Each of the 12 columns indicate a model. The three 'main models' studied were the Thompson model, the UCB model and a hybrid of both. Variants were then created by adding the ϵ-greedy parameter, the novelty bonus and a combination of both. All the parameters besides Q_0 and w were fitted to each horizon separately. Parameters: Q_0=prior mean (initial estimate of a bandits mean); σ_0=prior variance (uncertainty about Q_0); w=contribution of UCB vs Thompson; γ =information bonus; β=softmax inverse temperature; ϵ=ϵ-greedy parameter (stochasticity); η=novelty bonus. Model selection measures include the cross-validation held-out data likelihood averaged over subjects, mean (SD), as well as the subject count for which this model performed better over either 12 models or over the 3 best models.

		Mean (sd)			**Two-way repeated-measures ANOVA**			
	Horizon	Amisulpride	Placebo	Propranolol	**Main effect**		**Interaction**	
ϵ-greedy parameter	**short**	0.10 (0.10)	0.12 (0.08)	0.07 (0.08)	**D**	F(2, 54)=6.722, p=.002, η^2=.199	**DH**	F(2, 54)=1.305, p=.280, η^2=.046
	long	0.17 (0.14)	0.14 (0.10)	0.08 (0.06)	**H**	F(1, 54)=1.968, p=.166, η^2=.035	**HW**	F(1, 54)=6.08, p=.017, η^2=.101
Novelty bonus η	**short**	2.07 (0.98)	2.26 (1.37)	2.05 (1.16)	**D**	F(2, 54)=.249, p=.780, η^2=.009	**DH**	F(2, 54)=.03, p=.971, η^2=.001
	long	3.24 (1.19)	3.12 (1.63)	2.95 (1.70)	**H**	F(1, 54)=1.839, p=.181, η^2=.033	**HW**	F(1, 54)=8.416, p=.005, η^2=.135
Prior variance σ_0	**short**	1.18 (0.20)	1.12 (0.43)	1.25 (0.34)	**D**	F(2, 54)=.060, p=.942, η^2=.002	**DH**	F(2, 54)=2.162, p=.125, η^2=.074
	long	1.41 (0.61)	1.42 (0.59)	1.21 (0.44)	**H**	F(1, 54)=.129, p=.721, η^2=.002	**HW**	F(1, 54)=.022, p=.882, η^2<.001
Prior mean Q_0		3.22 (1.05)	3.20 (1.36)	3.44 (1.05)	**D**	F(2, 54)=.118, p=.889, η^2=.004		

Appendix 2 Table 6.

Table of statistics and fitted model parameters of Figure 5. The drug groups differed in ϵ-greedy parameter value (pairwise comparisons: placebo vs propranolol: t(40)=3.177, p=.002, d=.71; amisulpride vs placebo: t(38)=.251, p=.802, d=.057; amisulpride vs propranolol: t(38)=2.723, p=.009, d=.626). The main effect is either of drug group (D) or of horizon (H). The interaction is either drug-by-horizon (DH) or horizon-by-WASI (measure of IQ; HW).

	Horizon	Low exploration	High exploration	Additional parameters
Value-free random exploration	**short**	$\epsilon = 0.1$	$\epsilon = 0.2$	$\eta = 0$
	long	$\epsilon = 0.3$	$\epsilon = 0.4$	$\eta = 2$
Novelty exploration	**short**	$\eta = 0$	$\eta = 1$	$\epsilon = 0$
	long	$\eta = 2$	$\eta = 3$	$\epsilon = 0.2$
Thompson-sampling exploration	**short**	$\sigma_0 = 0.8$	$\sigma_0 = 1.2$	$\eta = 0, \epsilon = 0$
	long	$\sigma_0 = 1.6$	$\sigma_0 = 2$	$\eta = 2, \epsilon = 0.2$
UCB exploration	**short**	$\gamma = 0.1$	$\gamma = 0.3$	$\beta = 5, \epsilon = 0$
	long	$\gamma = 0.7$	$\gamma = 1.5$	$\beta = 1.5, \epsilon = 0.2$

Appendix 2 Table 7

Parameter values used for simulations on Figure 1- Figure supplement 3-5. Parameter values for high and low exploration were selected empirically from pilot and task data. Value-free random exploration and novelty exploration were simulated with an argmax decision function, which always selects the value with the highest expected value. For simulating the long (versus short) horizon condition, we assumed that not only the key value but also the other exploration strategies increased, as found in our experimental data. For each simulation $Q_0 = 5$ and unless otherwise stated, $\sigma_0 = 1.5$.

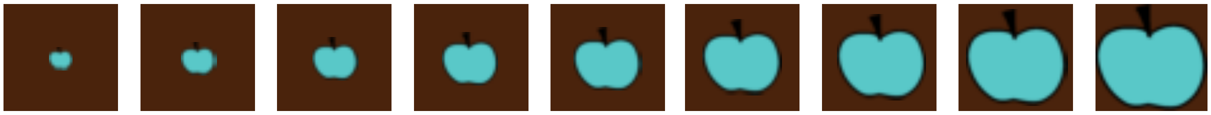


Figure 1 - Figure supplement 1

Visualisation of the 9 different sizes that the apples could take. The associated rewards went from 2 (small apple on the left) to 10 (big apple on the right).

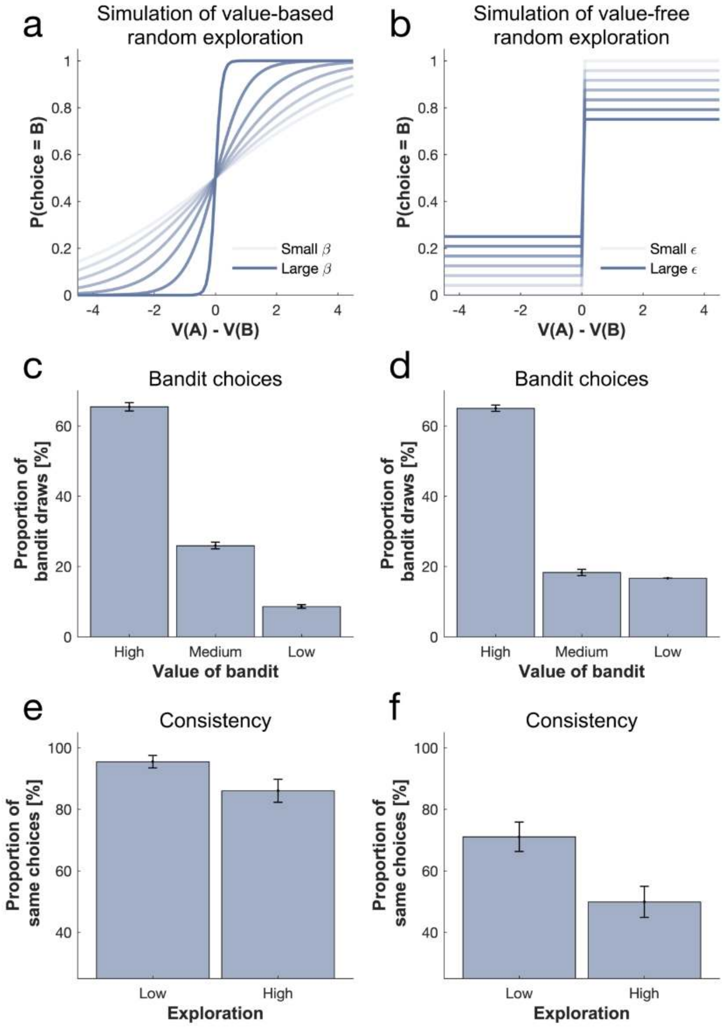


Figure 1 - Figure supplement 2

Comparison of value-based (softmax) and value-free (ϵ-greedy) random exploration. (a) Changing the softmax inverse temperature affects the slope of the sigmoid while changing the ϵ-greedy parameter (b) affects the compression of the sigmoid. Conceptually, in a softmax exploration mode, as each bandits' expected value is taken into account, (c) the 2nd best bandit (medium-value bandit) will be favoured over one with a lower value (low-value bandit) when injecting noise. In contrast, in an ϵ-greedy exploration mode, (d) bandits are explored equally often irrelevant of their expected value. Both simulations were performed on trials without novel bandit. When simulating on all trials we see that this also has a consequence on choice consistency, as (e) the 2nd best option will most probably be explored (i.e. choice is still more consistent) in a softmax exploration mode versus (f) equal probability of exploring any of the 2 non-optimal options in an ϵ-greedy exploration mode.

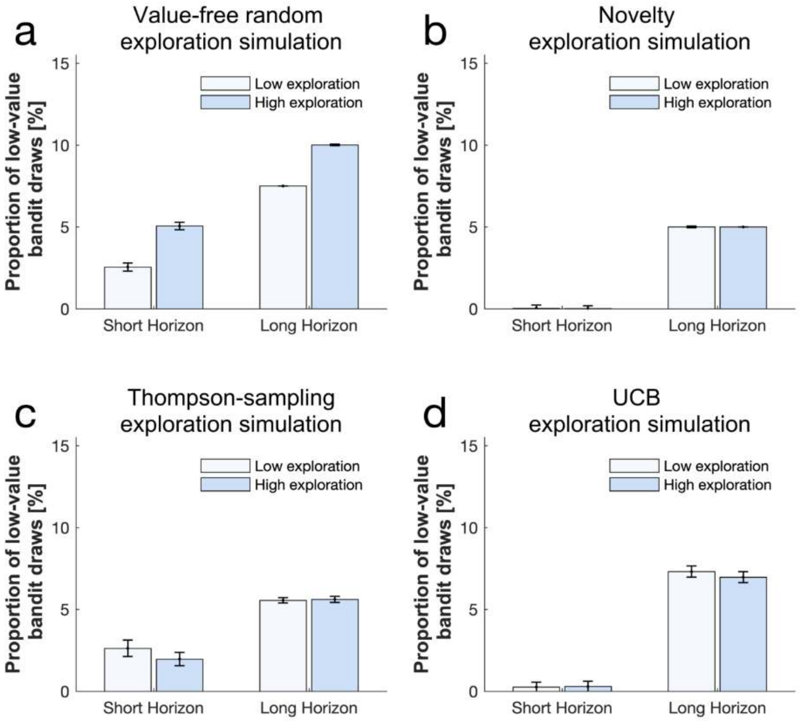


Figure 1 - Figure supplement 3

Simulating the effect of the different exploration strategies on the frequency of picking the low-value bandit shows that (a) a higher value-free random exploration increases the selection of the low-value bandit, whereas neither (b) a higher novelty exploration, (c) a higher Thompson-sampling exploration nor (d) a higher UCB exploration affected this frequency. For simulating the long (versus short) horizon condition, we assumed that not only the key value but also the other exploration strategies increased, as found in our experimental data (cf. Appendix 2 Table 7 for parameter values).

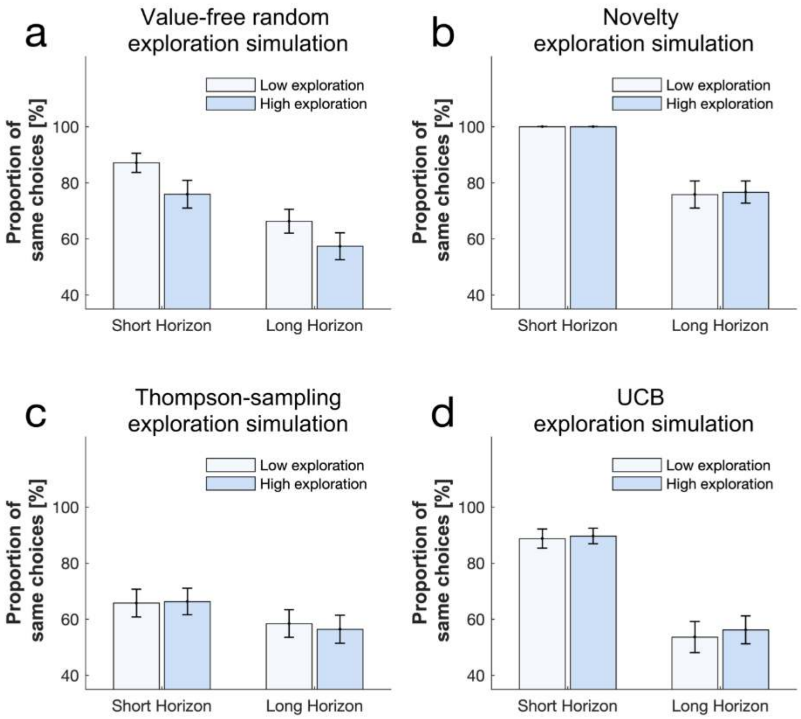


Figure 1 - Figure supplement 4

Simulating the effect of the different exploration strategies on choice consistency shows that (a) a higher value-free random exploration decreases the proportion of same choices, whereas neither (b) a higher novelty exploration, (c) a higher Thompson-sampling exploration nor (d) a higher UCB exploration affected this measure. For simulating the long (versus short) horizon condition, we assumed that not only the key value but also the other exploration strategies increased, as found in our experimental data (cf. Appendix 2 Table 7 for parameter values).

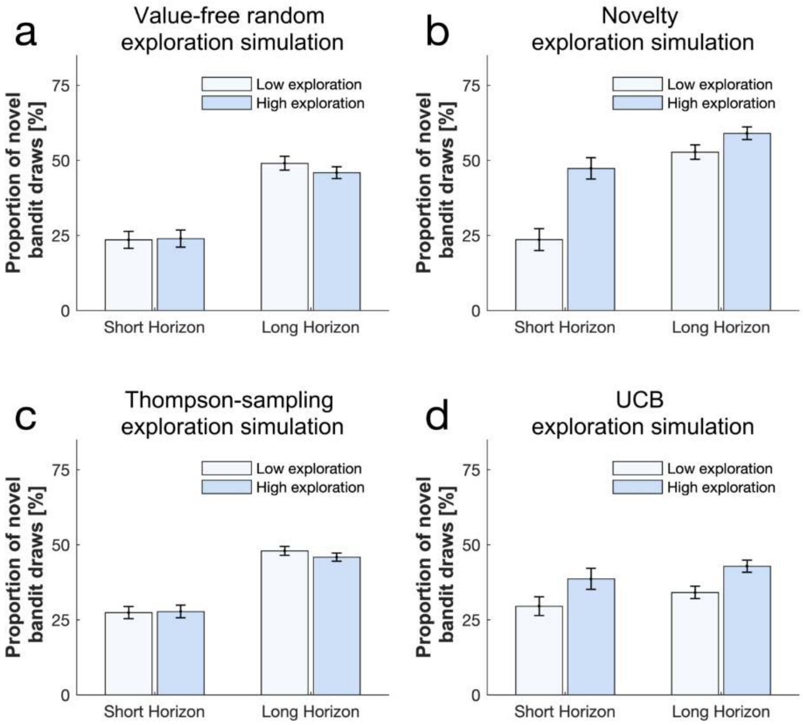


Figure 1 - Figure supplement 5

Simulating the effect of the different exploration strategies on the frequency of picking the novel bandit shows that (a) a higher value-free random exploration has little effect on the selection of the novel bandit, whereas (b) a higher novelty exploration increases this frequency. (c) A higher Thompson-sampling exploration had little effect and (d) a higher UCB exploration affected this frequency but to a lower extend than novelty exploration. For simulating the long (versus short) horizon condition, we assumed that not only the key value but also the other exploration strategies increased, as found in our experimental data (cf. Appendix 2 Table 7 for parameter values).

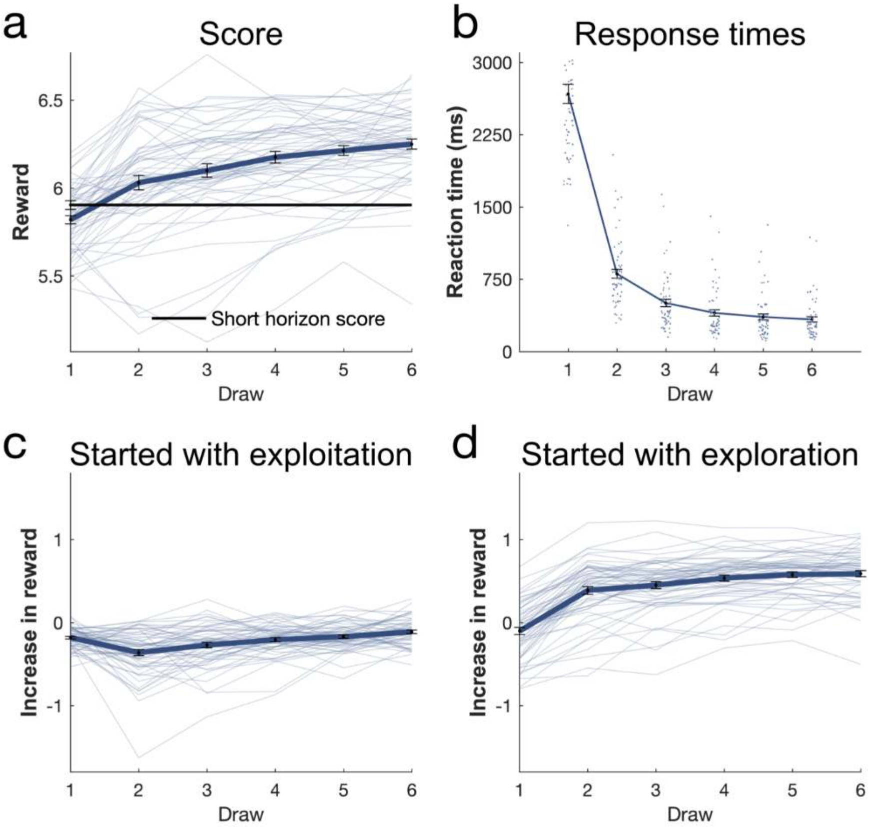


Figure 2 - Figure supplement 1

Further analysis of long horizon draws. (a) The first draw in the long horizon led to a lower reward than the short horizon, indicating more exploration, while the subsequent draws led to a higher reward indicating that this additional information helped making better decisions in the long run. (b) The first draws' response time was the highest and then decreased for each draw. Long horizon trials in which subjects started with (c) an exploitation draw (choose the bandit with the highest expected value) led to little increase in reward (y-axis: difference between obtained reward and highest reward of initial samples; linear regression slope coefficient: mean=0.118, sd=0.038), whereas trials in which they started with (d) an exploration draw led to an large increase in reward (linear regression slope coefficient: mean=0.028, sd=0.041). This larger increase in reward when starting by exploring (slope is higher: t(58)=-12.161, p<.001, d=-1.583) indicates that the information that was gained through exploration led to higher long-term outcomes. Data are shown as mean ± SEM and each dot represent one subject.

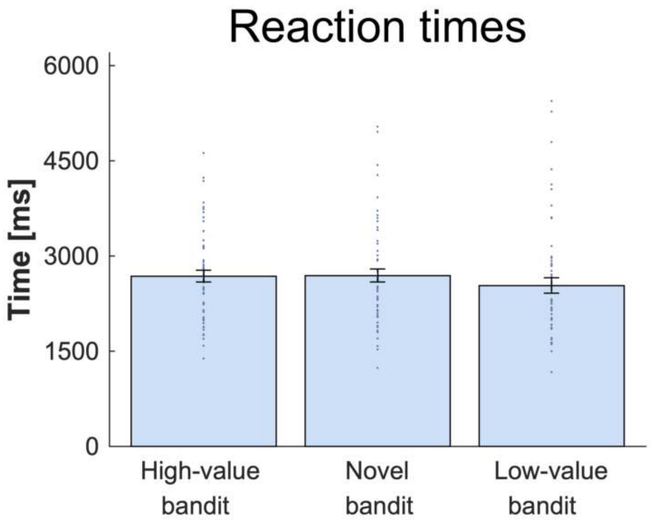


Figure 3 - Figure supplement 1

Response time analysis per bandit. There was no difference in RT depending which bandit was chosen. For details and statistics cf. Appendix 1.

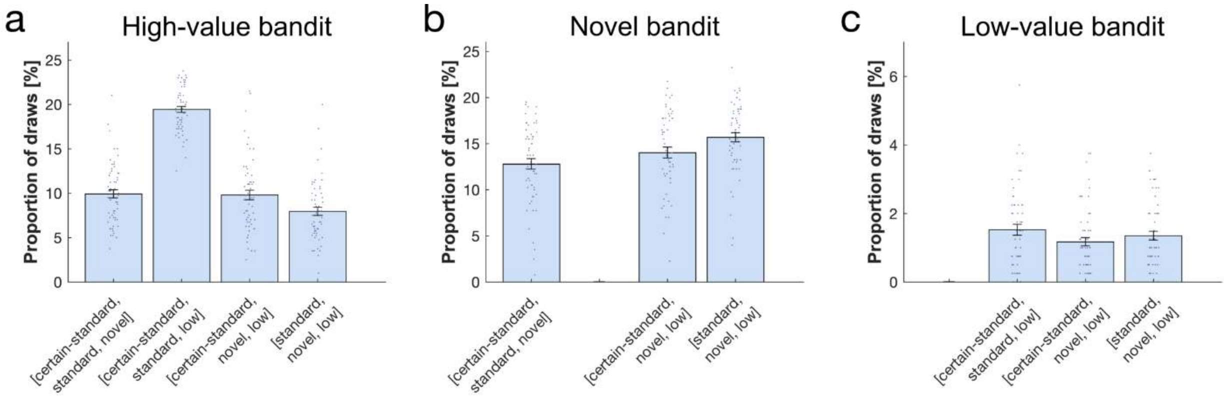


Figure 3 - Figure supplement 2

Proportion of draws per bandit combination (x-axis). (a) The high-value bandit was picked more when there was no novel bandit, and less when the high-value bandit was less certain. (b) The novel bandit was picked the most when the high-value bandit was less certain, then when the high-value bandit was more certain, and it was picked the least when both certain and certain standard bandits were present. (c) The low-value bandit was picked less when the high-value bandit was more certain. For statistics see Appendix 1.

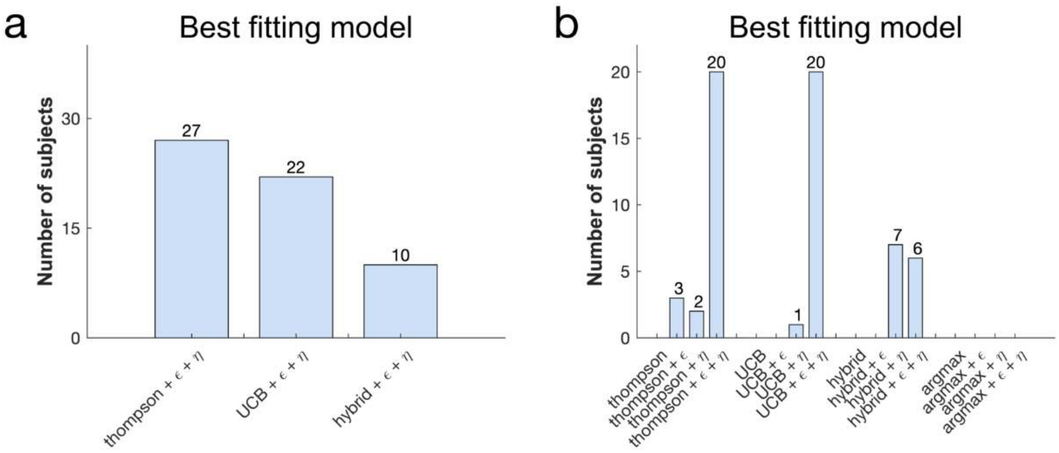


Figure 4 - Figure supplement 1

Model comparison: further evaluations. (a) The winning model at the group level (the Thompson model with both ϵ and η) was also the one that accounted best for the largest number of subjects. (b) The Thompson+ϵ+η model and the UCB+ϵ+η are equally first in subject count when comparing all models, the Thompson+ϵ+η model is therefore still the winning model as it has the highest average likelihood of held-out data.

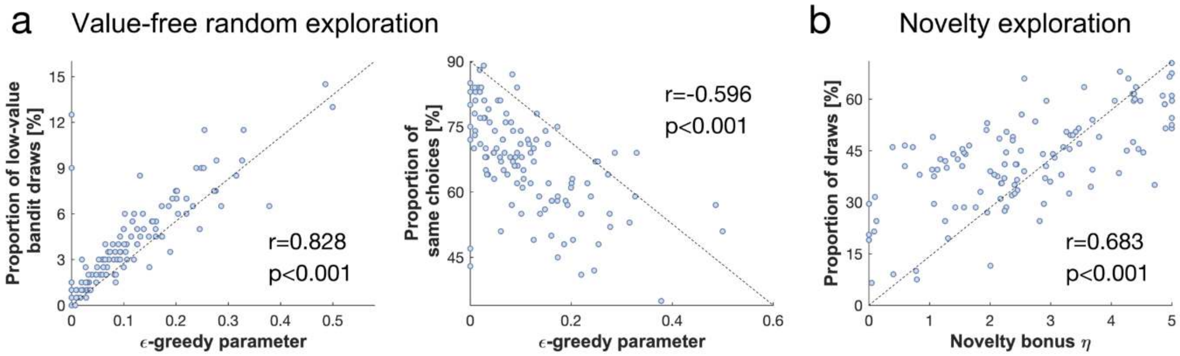


Figure 4 - Figure supplement 2

Correlations between model parameters and behaviour. The behavioural indicators of (a) value-free random exploration (left panel: draws from the low-value bandit; right panel: consistency) correlated with the ϵ-greedy parameter values, and of (b) novelty exploration (draws from the novel bandit) correlated with the novelty bonus η.

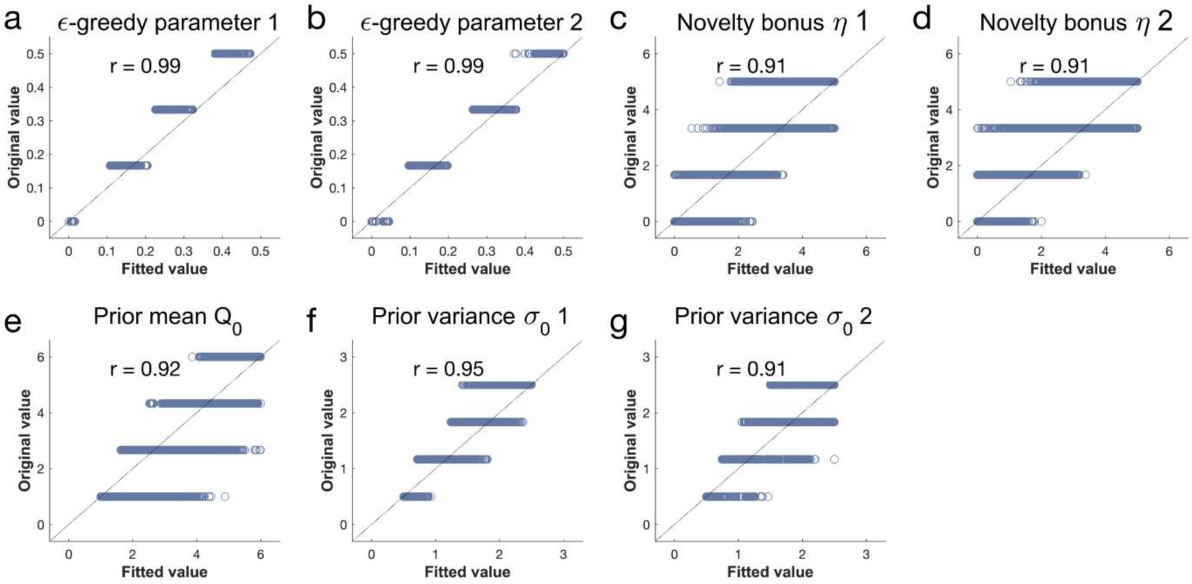


Figure 4 - Figure supplement 3

Parameter recovery analysis details. For each of the 7 parameters of the winning model, we took 4 values, equally spread within the parameter range. We simulated behaviour using every combination ($4^7 = 16384$), fitted the model and analysed how well the generative parameters (original values) correlated with the recovered ones (fitted parameters). Pearson correlation coefficient = r. Each dot represents one simulation.

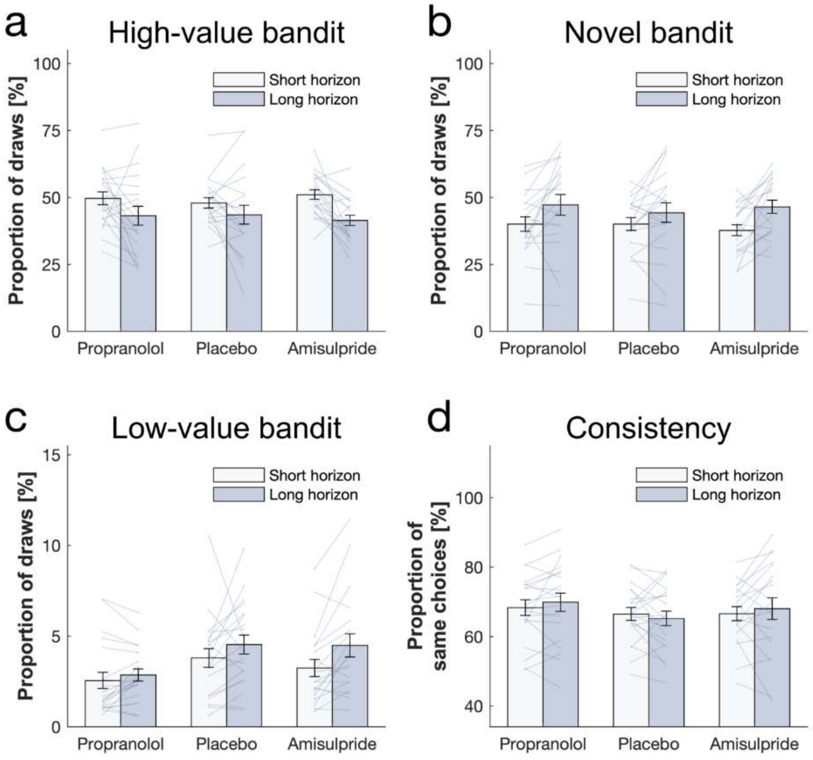
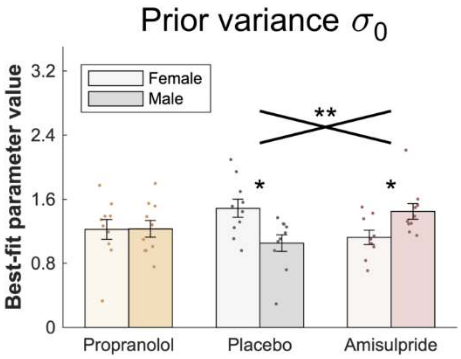


Figure 5 - Figure supplement 1

Simulated behaviour. We used each subjects' fitted parameters to simulate behaviour (blue diamonds; N_{trials}=4000) and superposed them to the real behaviour measures (N_{trials}=400) measures. Data are shown as mean ± SEM and each dot/line represent one agent.



1347

1348 **Figure 5 – Figure supplement 2**

1349 Gender effect on prior variance parameter. Mean values (across horizon conditions) of σ_0 were larger for female subjects, whereas in the amisulpride group, they were larger for male subjects. Data are shown as mean ± SEM and each dot represent one subject.