Noradrenaline modulates tabula-rasa exploration

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: 45
Number of Figures: 5
Number of Tables: 0
Abstract: 124 words
Introduction: 764 words
Discussion: 1520 words

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

Abstract

An exploration-exploitation trade-off, the arbitration between sampling lesser-known alternatives against harvesting a known rich option, is thought to be solved by humans using computationally demanding exploration algorithms. Given known limitations in human cognitive resources, we hypothesised that additional energy efficient strategies are deployed. We examine such heuristics in choice behaviour and using computational modelling show this involves a tabula-rasa exploration that ignores all prior knowledge and a novelty exploration that only targets novel options. In a double-blind, placebo-controlled drug study, assessing contributions of dopamine (400mg amisulpride) and noradrenaline (40mg propranolol), we show that tabula-rasa exploration is attenuated when noradrenaline, but not dopamine, is blocked. Our findings demonstrate that humans deploy distinct computationally efficient exploration strategies and that one of these 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 your exploration strategy, you may end up with a highly disappointing ice cream encounter, or a life-changing gustatory epiphany.

A common procedure when studying complex decision making, for example an exploration-exploitation trade-off, is to avail of computational algorithms developed in artificial intelligence and test whether their signatures are evident in human behaviour. This has revealed humans use strategies that reflect implementation of computationally demanding exploration algorithms (*1*, *2*). One such strategy, directed exploration, involves awarding an 'information bonus' to choice options and this bonus is scaled with uncertainty. This is captured in algorithms such as the Upper Confidence Bound (*3*, *4*) (UCB) and leads to the 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 in the decision process. This can be realised using a fixed parameter as source of stochasticity, such as a softmax temperature parameter (*6*, *7*), which is usually 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 the uncertainty about choice options. This uncertainty dependent exploration is

58 essentially a more sophisticated version of a softmax. By accounting for stochasticity when
59 comparing the choice options' expected values, these exploration strategies increase the likelihood
60 of choosing options slightly inferior to the one believed to be the best (e.g. the Toblerone ice-
61 cream).

62 The above processes are computationally demanding, especially in real-life multiple-
63 alternative decision problems, and may exceed one's available cognitive resources (*6*, *9*, *10*).
64 Indeed, the computational algorithms that have been suggested (i.e. UCB and Thompson sampling)
65 require computation and representation of the expected value of all possible choice options plus
66 their respective uncertainties (or variance). Here, we propose and examine the explanatory power
67 of two additional computationally less costly forms of exploration, namely tabula-rasa and novelty
68 exploration.

69 Computationally the most efficient way to explore is to ignore all prior information and to
70 choose entirely randomly, de facto assigning the same probability to all options. Such 'tabula-rasa
71 exploration' forgoes costly computations and is a reflection of what is known in reinforcement
72 learning as an ϵ-greedy algorithmic strategy (*11*). The computational efficiency, however, comes
73 at a cost of sub-optimality due to occasional selection of options of low expected value (e.g. the
74 repulsive spinach ice cream). Despite its sub-optimality, tabula-rasa exploration has
75 neurobiological plausibility. Amongst other roles (such as memory (*12*), or energisation of
76 behaviour (*13*, *14*)), the neurotransmitter noradrenaline has been ascribed the function of reflecting
77 uncertainty (*15–17*) and as a 'reset button' that interrupts ongoing information processing (*18–*
78 *20*). Prior experimental work in rats shows that boosting noradrenaline leads to more tabula-rasa-
79 like behaviour (*21*). Here, we examine whether we can identify signatures of tabula-rasa

exploration in human decision making and whether manipulating noradrenaline impacts its expression.

An alternative computationally efficient exploration heuristic is to simply choose an option not encountered previously. Humans often show novelty seeking (*22–25*), and this strategy can also be used in exploration and can be implemented by a low-cost version of the UCB algorithm (i.e. does not rely on precise uncertainty estimates), in which a novelty bonus (*26*) is added only to previously unseen choice options. The neurotransmitter dopamine has been implicated not only in exploration in general (*27*), but also in signalling such types of novelty bonuses, where evidence indicates its role in processing and exploring novel and salient states (*23*, *28–31*).

In the current study, we examine the contributions of tabula-rasa and novelty exploration in human choice behaviour. We developed a novel exploration task and computational models to probe 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 tabula-rasa 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 tabula-rasa 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. In contrast, UCB also assigns a bonus to choice options that have existing, but less detailed, information. Thus, 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). Novelty exploration predicts that a novel option will be chosen disproportionally more often (Figure 1f).

Tabula-rasa exploration 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. Such known, poor value, options are not favoured by other exploration strategies (Figure 1e). A second prediction is that the choice consistency is substantially affected by tabula-rasa exploration. Given that tabula-rasa exploration splits its choice probability equally (i.e. 33.3% of choosing any bandit out of the three displayed), it is less likely to choose again the same bandit even when facing 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).

We generated bandits from four different generative groups (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 tabula-rasa exploration strategy alone. The last bandit was a ‘novel bandit’ for which no initial sample was shown, appealing to a novelty exploration strategy. To assess choice consistency, all trials were repeated once. To avoid that some exploration strategies overshadow others, only three of the four different bandit types were available to choose from on each trial. Lastly, to assess whether subjects’ behaviour was meaningful in terms of 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*, *32*).

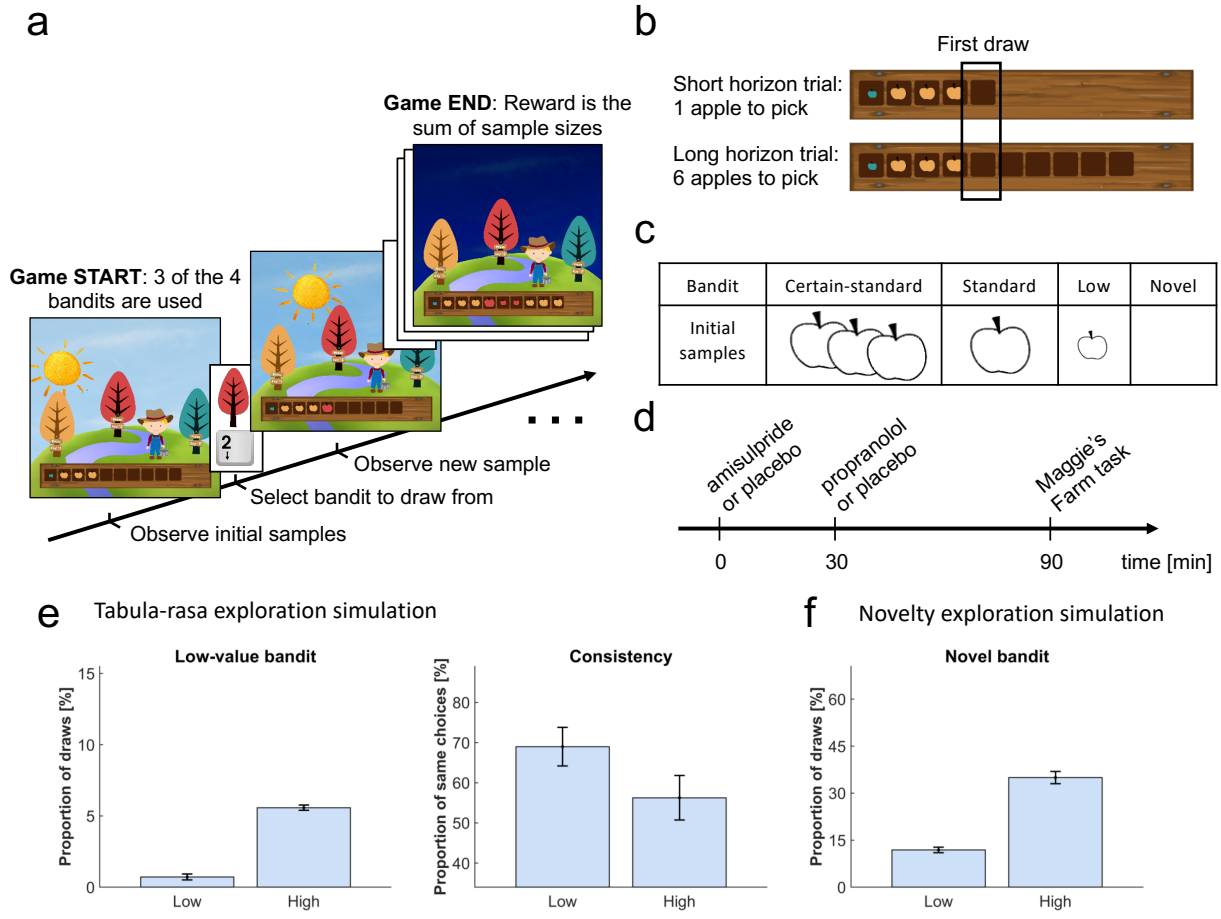

Figure 1. Study design. In the Maggie's farm task, subjects had to choose from three bandits (depicted as trees) to maximise an outcome (sum of reward). The rewards (apple size) of each bandit followed a normal distribution with a fixed sampling variance. (a) At the beginning of each trial, subjects 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. (c) In each trial, three bandits were displayed, selected from four possible bandits, with different generative groups 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. (d) Prior to the task, subjects were administered different drugs: 400mg amisulpride for the dopamine group, 40mg propranolol for the noradrenaline group, 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 tabula-rasa behaviour in this task shows that in a high tabula-rasa 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.

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: dopamine, noradrenaline or placebo. The noradrenaline group was administered 40mg of the β-adrenoceptor antagonist propranolol, while the dopamine group was administered 400mg of the D2/D3 antagonist amisulpride. 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 (WASI (*33*)) and mood as drug groups differed marginally in those measures (cf. Appendix 2 Table 1). Similar results were obtained when not correcting for these 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). As expected, subjects explored more in the long compared to the short horizon (repeated-measures ANOVA for the expected value: $F(1, 56)=19.457$, $p<.001$, $\eta 2=.258$; Figure 2a). When 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) we found that subjects

(especially those with higher IQ) sampled from it more in the short compared to the long horizon, (horizon main effect: $F(1, 54)=3.909$, $p=.053$, $\eta 2=.068$; WASI-by-horizon interaction: $F(1,54)=13.304$, $p=.001$, $\eta 2=.198$; Figure 3a), confirming a reduction in exploitation when this information could be subsequently used. Interestingly, this frequency seemed to be marginally higher in the dopamine 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 ($F(1, 56)=23.922$, $p<.001$, $\eta 2=.299$; 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 ($F(1, 56)=58.78$, $p<.001$, $\eta 2=.512$; 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 ($F(1, 56)=103.759$, $p<.001$, $\eta 2=.649$; 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.

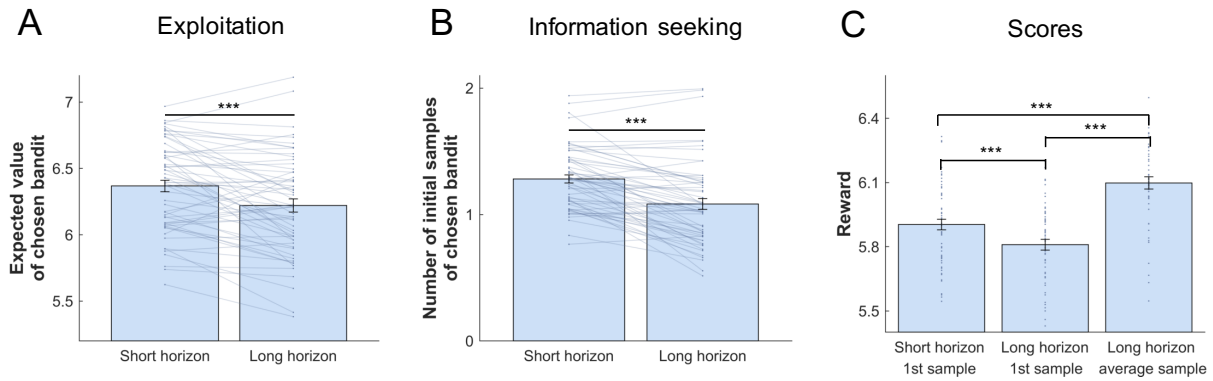

Figure 2. Benefits of exploration. To investigate the effect of information on performance we collapsed subjects over all three treatment groups. (a) The first bandit subjects chose as a function of its expected value (average of its initial samples). Subjects chose bandits with a lower expected value (i.e. they exploited less) in the long horizon compared to the short horizon. (b) The first bandit subjects chose as a function of the number of samples that were initially revealed. Subjects chose less known (i.e. more informative) bandits 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. *** =p<.001. Data are shown as mean ± SEM and each dot/line represents a subject.

Subjects demonstrate tabula-rasa behaviour

Tabula-rasa 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 and thought through strategies), the probability of choosing bandits with a lower expected value (here the low-value bandit, Fig 1e) will be higher. We investigated whether such behavioural signatures were increased in the long horizon condition (i.e. when exploration is useful), and we found a significant main effect of horizon ($F(1, 54)=4.069$, $p=.049$, $\eta 2=.07$; Figure 3c). This demonstrates that tabula-rasa exploration is utilised more when exploration is beneficial.

Tabula-rasa behaviour is modulated by noradrenaline function

When we tested whether tabula-rasa exploration was sensitive to neuromodulatory influences, we found a difference in how often drug groups sampled from the low-value option

(drug main effect: $F(2, 54)=7.003$, $p=.002$, $\eta 2=.206$; drug-by-horizon interaction: $F(2, 54)=2.154$, $p=.126$, $\eta 2=.074$; Figure 3c). This was driven by the noradrenaline group choosing the low-value option significantly less often than the other two groups (placebo vs noradrenaline: $t(40)=2.923$, $p=.005$, $d=.654$; dopamine vs placebo: $t(38)=-0.587$, $p=.559$, $d=.133$; dopamine vs noradrenaline: $t(38)=2.171$, $p=.034$, $d=.496$). These findings demonstrate that a key feature of tabula-rasa exploration, the frequency of choosing low-value bandits, is sensitive to influences from noradrenaline.

To further examine drug effects on tabula-rasa exploration, we assessed a second prediction, namely choice consistency. Because tabula-rasa exploration ignores all prior information and chooses randomly, it should result in a decreased choice consistency when presented identical choice options (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 the consistency with which drug groups sampled from different option (drug main effect: $F(2, 54)=7.154$, $p=.002$, $\eta 2=.209$; horizon main effect: $F(1, 54)=1.333$, $p=.253$, $\eta 2=.024$; drug-by-horizon interaction: $F(2, 54)=3.352$, $p=.042$, $\eta 2=.11$; Figure 3d), driven by the fact that the noradrenaline group chose significantly more consistently than the other two groups (pairwise comparisons: placebo vs noradrenaline: $t(40)=-3.525$, $p=.001$, $d=.788$; dopamine vs placebo: $t(38)=1.107$, $p=.272$, $d=.251$; dopamine vs noradrenaline: $t(38)=-2.267$, $p=.026$, $d=.514$). Taken together, these results indicate that tabula-rasa exploration depends critically on noradrenaline functioning, such that an attenuation of noradrenaline leads to a reduction in tabula-rasa exploration.

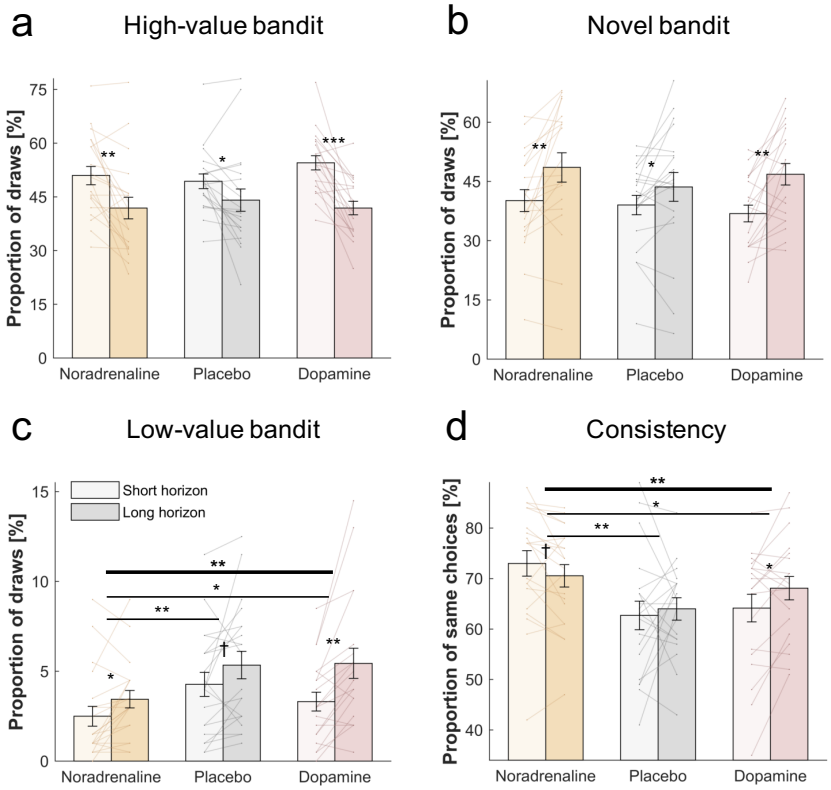

Figure 3. Behavioural horizon and drug effects. Choice patterns in the first draw for each horizon and drug group (noradrenaline, placebo and dopamine). Subjects from all three groups sampled from the (a) high-value bandit (i.e. bandit with the highest average reward of initial samples) more in the short horizon compared to the long horizon (indicating reduced exploitation) and from the (b) novel bandit more in the long horizon compared to the short horizon (novelty exploration). (c) All groups sampled from the low-value bandit more in the long horizon compared to the short horizon (tabula-rasa exploration), but subjects in the noradrenaline group sampled less from it, and (d) were more consistent in their choices overall, indicating that noradrenaline blockade reduces tabula-rasa exploration. Horizontal bars represent rm-ANOVA (thick) and pairwise comparisons (thin). † =p<.10, * =p<.05, ** =p<.01, *** =p<.001. Data are shown as mean ± SEM and each dot/line represent one participant.

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 1g). We found a significant main effect of horizon ($F(1, 54)=5.593$, $p=.022$, $\eta2=.094$; WASI-by-horizon interaction: $F(1, 54) =13.897$, $p<.001$, $\eta2=.205$; Figure 3b) indicating that subjects explored the novel bandit significantly more often in the long horizon

256 condition, and this was particularly strong for subjects with a higher IQ. We next assessed whether
257 novelty exploration was sensitive to our drug manipulation, but found no drug effects on the novel
258 bandit ($F(2, 54)=1.498$, $p=.233$, $\eta 2=.053$; drug-by-horizon interaction: $F(2, 54)=.542$, $p=.584$,
259 $\eta 2=.02$; Figure 3c), indicating that an attenuation of dopamine or noradrenaline function does not
260 impact novelty exploration in this task.

261 *Subjects combine computationally demanding strategies and exploration heuristics*

262 To examine the contributions of different exploration strategies to choice behaviour, we
263 fitted a set of computational models to subjects' behaviour, building on models developed in
264 previous studies (*1*). In particular, we compared models incorporating the UCB algorithm, softmax
265 decision function, Thompson sampling, an ϵ-greedy algorithm and the novelty bonus (cf. Materials
266 and Methods). Essentially, each model makes different exploration predictions. In the Thompson
267 model, Thompson sampling (*8*, *34*) leads to an uncertainty-driven exploration, where both
268 expected value and uncertainty contribute to choice. In this model higher uncertainty leads to more
269 exploration such that instead of selecting a bandit with the highest mean, bandits are chosen
270 relative to how often a random sample would yield the highest outcome, thus accounting for
271 uncertainty (*2*). In the UCB model, the UCB algorithm (*3*, *4*), capturing directed exploration,
272 predicts that each bandit is chosen according to a mixture of expected value and an additional
273 expected information gain (*2*). This is realised by adding a bonus to the expected value of each
274 option, proportional to how informative it would be to select this option (i.e. the higher the
275 uncertainty in the options' value, the higher the information gain). This computation is then passed
276 through a softmax decision model, capturing random exploration. Novelty exploration is a
277 simplified version of the information bonus in the UCB algorithm, which only applies to entirely
278 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 tabula-rasa ϵ-greedy algorithm selects any bandit ϵ % of the time, irrespective of the prior information of this bandit.

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 Thompson sampling with tabula-rasa 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 the UCB model and the Thompson model explained the data better than those models alone, but this was no longer the case when accounting for novelty and tabula-rasa exploration (Figure 4a). The winning model further revealed that all parameter estimates could be accurately recovered (Figure 4b; Figure 4 – Figure supplement 3). In summary, the model selection suggests that in our experiment subjects used a mixture of computationally demanding (i.e. Thompson sampling) and heuristic exploration strategies (i.e. tabula-rasa and novelty exploration).

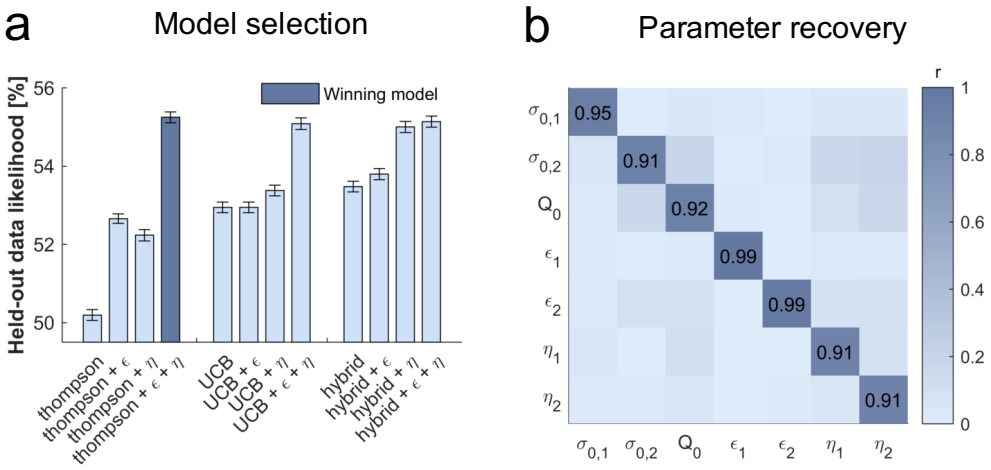

Figure 4. Subjects use a mixture of exploration strategies. (a) A 10-fold cross-validation of the likelihood of held-out data was used for model selection (chance level =33.3%; cf. Figure 4 – Figure supplement 1 for model selection at the individual level). The Thompson model with both the tabula-rasa 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 (cf. Figure 4 – Figure supplement 2 for correlations between behaviour and model parameters); σ_0 is the prior variance and Q_0 is the prior mean (cf. Figure 4 – Figure supplement 3 for parameter recovery correlation plots and Materials and Methods for model details); 1 stands for short horizon-, and 2 for long horizon-specific parameters.

Noradrenaline controls tabula-rasa 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 2 for exact values). First, we examined the ϵ-greedy parameter that captures the contribution of tabula-rasa exploration to choice behaviour. We assessed how this tabula-rasa 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 noradrenaline: t(40)=3.177, p=.002, d=.71;

dopamine vs placebo: t(38)=.251, p=.802, d=.057; dopamine vs noradrenaline: t(38)=2.723, p=.009, d=.626).

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 tabula-rasa exploration.

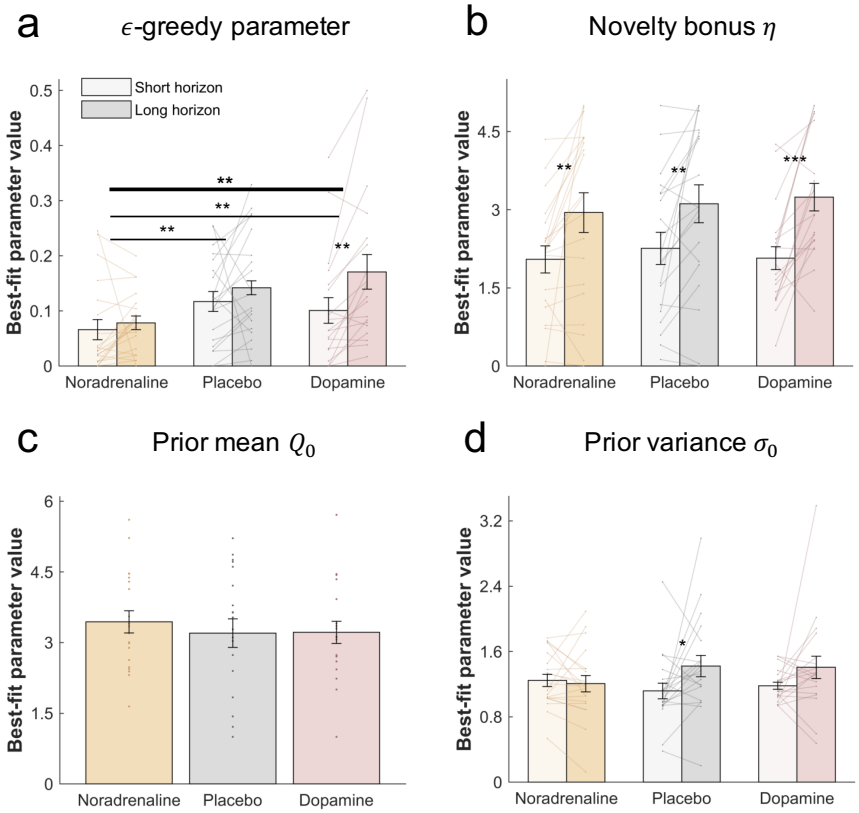

Figure 5. Drug effects on model parameters. The winning model's parameters were fitted to each subject's first draw (cf. Figure 5 – Figure supplement 1 for model simulations). (a) All groups had higher values of ϵ (tabula-rasa exploration) in the long compared to the short horizon. Notably, subjects in the noradrenaline group had lower values of ϵ overall, indicating that attenuation of noradrenaline functioning reduces tabula-rasa 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 (cf. Figure 5 – Figure supplement 2 for gender effects). * =p<.05, ** =p<.01, *** =p<.001.

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, η2=.009; drug-by-horizon interaction: F(2, 54)=.03, p=.971, η2=.001). There was also a close alignment between model-based and model-free analyses (correlation between the novelty bonus η with draws from the novel bandit: $R_{Pearson}$=.683,

339 p<.001; Figure 4 – Figure supplement 2), and we found a similarly increased novelty bonus effect
340 in the long horizon in subjects with a higher IQ (horizon main effect: F(1, 54)=1.839, p=.181,
341 η2=.033; WASI-by-horizon interaction: F(1, 54) =8.416, p=.005, η2=.135; Figure 5b).

342 When analysing the additional model parameter, we found that subjects had similar prior
343 beliefs about bandits, given by the initial estimate of a bandit's mean (prior mean Q_0: F(2,
344 54)=.118, p=.889, η2=.004; Figure 5c) and their uncertainty about it (prior variance σ_0: horizon
345 main effect: F(1, 54)=.129, p=.721, η2=.002; drug main effect: F(2, 54)=.06, p=.942, η2=.002;
346 drug-by-horizon interaction: F(2, 54)=2.162, p=.125, η2=.074; WASI-by-horizon interaction: F(1,
347 54)=.022, p=.882, pes<.001; Figure 5d). Interestingly, our dopamine manipulation seemed to
348 affect this uncertainty in a gender-specific manner, with female participants having larger values
349 of σ_0 compared to males in the placebo group, and with the opposite being true in the dopamine
350 group (cf. Appendix 1). Taken together, these findings show that tabula-rasa exploration was most
351 sensitive to our drug manipulations.

Discussion

Solving the exploration-exploitation problem is not trivial, and it has been suggested that humans solve it using computationally demanding exploration strategies (*1*, *2*). These take the uncertainty (variance) as well as the expected reward (mean) of each choice option into account, a demand of computational resources which might not always be ensured (*10*). Here, we demonstrate that two additional, computationally less resource-hungry exploration heuristics are at play during human decision-making, namely tabula-rasa and novelty exploration.

By assigning an intrinsic value (novelty bonus (*26*)) to an option not encountered before (*35*), a novelty bonus can be seen as an efficient simplification of demanding algorithms, such as UCB (*3*, *4*). It is interesting to note that the winning model did not include UCB, but novelty exploration. This indicates that humans may 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, tabula-rasa exploration, also plays a role in our task, a strategy that also requires minimal computational resources. Even though less optimal, its simplicity and neural plausibility render it a viable strategy. In our study, converging behavioural and modelling indicators demonstrate that both tabula-rasa and novelty exploration were deployed in a goal-directed manner, coupled with increased levels of exploration when this was strategically useful. Exploration was captured similarly 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 for extended exploration (Wilson et al., 2014), which we indeed find in our study. Although tabula-rasa exploration can

375 comprise influences of attentional lapses or impulsive motor responses, the difference between
376 horizon conditions cancels them out. Importantly, our results demonstrate that this tabula-rasa
377 exploration is at play, even when accounting for more sophisticated forms of stochasticity (*1*, *7*),
378 whether fixed or uncertainty-driven, captured by our more complex models.

379 It is still unclear how exploration strategies are implemented neurobiologically.
380 Noradrenaline inputs, arising from the locus coeruleus (*36*) (LC) are thought to modulate
381 exploration (*2*, *37*, *38*), though empirical data on its precise mechanisms and means of action
382 remains limited. In this study, we found that noradrenaline impacted tabula-rasa exploration,
383 suggesting that noradrenaline may disrupt ongoing valuation processes and discard prior
384 information. This is consistent with findings in rodents wherein enhanced anterior cingulate
385 noradrenaline release led to more random behaviour (*21*) and with pharmacological findings in
386 monkeys that show enhanced choice consistency after reducing LC noradrenaline firing rates (*39*).
387 Aside from invasive methods and pharmacological manipulations, pupil diameter is used as an
388 indirect marker of noradrenaline activity (*40*), although the link between the two it not always
389 straightforward (*41*). Our current findings are also in line with previous human studies that showed
390 an association between this indirect marker and exploration, but did not dissociate between the
391 different potential exploration strategies that subjects can deploy (*42*). Future studies that replicate
392 our results could complement our findings by including indirect measures of noradrenaline
393 activity, for example pupillometry, to examine a potential link between natural variations in
394 noradrenaline levels and a propensity to tabula-rasa exploration.

395 The LC has two known modes of synaptic signalling (*36*), tonic and phasic, and these are
396 thought to have complementary roles (*20*). Phasic noradrenaline is thought to act as a reset button
397 (*20*), rendering an agent agnostic to all previously accumulated information, a de facto signature

398 of tabula-rasa exploration. Tonic noradrenaline has been associated, although not consistently (*43*),
399 with increased exploration (*37*, *44*), decision noise in rats (*45*) and more specifically with random
400 as opposed to directed exploration strategies (*32*). This later study unexpectedly found that
401 boosting noradrenaline decreased (rather than increased) random exploration, which the authors
402 speculated was due to an interplay with phasic signalling. Importantly, the drug used in that study
403 also, albeit to a lesser extent, affects dopamine function making it difficult to assign a precise
404 function to noradrenaline. This latter influenced our decision to opt for drugs with high specificity
405 for either dopamine or noradrenaline (*46*), thus enabling us to reveal highly specific effect on
406 tabula-rasa exploration. Although the contributions of tonic and phasic noradrenaline signalling
407 cannot be disentangled in our study, our findings do align with theoretical accounts and non-
408 primate animal findings, indicating that phasic noradrenaline promotes tabula-rasa exploration.

409 Aside from this 'reset signal' role, noradrenaline has been assigned other functions. It has
410 been linked to memory (*12*, *47*, *48*), which is not at play here since all the information is always
411 visible on the screen, and is thought to be involved in volatility and uncertainty estimation (*15–*
412 *17*). In particular it has more recently been linked to the energisation of behaviour. Non-human
413 primates studies have demonstrated a higher LC activation for high effort choices, suggesting that
414 noradrenaline release facilitates energy mobilisation (*13*). Theoretical models suggest that the LC
415 is involved in the control of effort exertion: It is thought to help trading off between effortful
416 actions leading to large rewards and "effortless" actions leading to small rewards by modulating
417 "raw" reward values by the required effort (*14*). Our task can be interpreted as such a trade-off:
418 complex exploration strategies are effortful but optimal in terms of reward gain, while tabula-rasa
419 exploration requires little effort but occasionally leads to a low reward. Applying this model, a
420 noradrenaline boost could optimise cognitive effort allocation for high reward gain (*14*), thereby

421 facilitating complex exploration strategies compared to tabula-rasa exploration. In such a
422 framework however, blocking noradrenaline release should decrease usage of complex exploration
423 strategies, leading to an increase of tabula-rasa exploration which is the opposite of what we
424 observed in our data. Another interpretation of this effort-facilitation model of noradrenaline could
425 be that a noradrenaline boost would help overcoming the cost, i.e. the lack of immediate reward
426 when selecting the low-value bandit, essentially giving a significative increase to the value of
427 information gain. In line with our results, a decrease in noradrenaline would interrupt this increase
428 in valuation, thereby removing the incentive to choose the low-value option. This theory however
429 is currently limited by the fact that there is no direct empirical evidence of noradrenaline boosting
430 valuations, and that it would also lead to an increased valuation, and therefore selection, of the
431 novel bandit which we did not observe in our data.

432 Dopamine has been ascribed multiple functions besides reward learning (*49*), such as
433 novelty seeking (*30*, *50*, *51*) or exploration in general (*27*). In fact, studies have demonstrated that
434 there are different types of dopaminergic neurons in the ventral tegmental area, and that some of
435 them are involved in non-reward signals, such as saliency and novelty (*28*). This suggests a role
436 in novelty exploration. Moreover, dopamine has also been suggested to be important in the
437 exploration-exploitation arbitration (*52–54*), although its role remains rather unclear with reported
438 dopaminergic effects on random exploration (*55*), on directed exploration (*29*, *56*), or no effects
439 at all (*57*). Our results did not demonstrate any main effect of dopamine manipulation on
440 exploration strategies, even though blocking dopamine led to a trend level increase in exploitation
441 (cf. Appendix 1). We believe that this is unlikely to reflect an ineffective drug dose as previous
442 studies have found neurocognitive effects with the same dose (*41*, *46*, *58*, *59*). One possible reason
443 for these light findings is that our dopaminergic blockade targets D2/D3 receptors rather than D1

444 receptors, a limitation due to the fact that there are no specific D1 receptor blockers available in
445 humans. An expectation of greater D1 involvement arises out of theoretical models (*60*) and a
446 prefrontal hypothesis of exploration (*56*). Interestingly, we did observe a minor gender-specific
447 differential drug effect on the subjects' uncertainty about an expected reward, with women being
448 more uncertain than men in the placebo setting, but more certain than them in the dopamine
449 blockade setting (cf. Appendix 1). This might be meaningful as other studies using the same drug
450 have also found behavioural gender-specific drug effect (*61*). Upcoming, novel drugs (*62*) might
451 be able help unravel a D1 contribution to different forms of exploration. Additionally, future
452 studies could use approved D2/D3 agonists (e.g. ropinirole) in a similar design to probe further
453 whether enhancing dopamine leads to a general increase in exploration.

454 In conclusion, humans supplement computationally expensive exploration strategies with
455 less resource demanding exploration heuristics, namely tabula-rasa and novelty exploration. Our
456 finding that noradrenaline specifically influences tabula-rasa exploration demonstrates that distinct
457 exploration strategies are under specific neuromodulator influence. Importantly, our data does not
458 allow to draw final conclusions about when and how the tabula-rasa exploration process (and by
459 extent its behavioural signatures) takes place: It could happen on a high-level (prefrontal)
460 mechanism or it could be implemented at a low-level (motor) degree. Future studies should
461 investigate this further. Our current findings can enable a richer understanding of disorders of
462 exploration, such as attention-deficit/hyperactivity disorder (*63*, *64*) and how aberrant
463 catecholamine function leads to its core behavioural impairments.

464 **Materials and Methods**

465 *Subjects*

466 Sixty healthy volunteers aged 18 to 35 (mean =23.22, SD =3.615) participated in a double-
467 blind, placebo-controlled, between-subjects study. The sample size was determined using power
468 calculation taking effect sizes from our prior studies that used the same drug manipulations (*41*,
469 *46*, *65*). Each subject was randomly allocated to one of three drug groups, controlling for an equal
470 gender balance across all groups (cf. Appendix 1). Candidate subjects with a history of
471 neurological or psychiatric disorders, current health issues, regular medications (except
472 contraceptives), or prior allergic reactions to drugs were excluded from the study. The groups
473 consisted of 20 subjects each matched (cf. Appendix 2 Table 1) for gender and age. Subjects
474 differed in mood (PANAS (*66*)) and marginally in intellectual abilities (WASI abbreviated version
475 (*33*)), which is why we control for these potential confounders in our analyses (cf. Appendix 1 for
476 uncorrected results). Subjects were reimbursed for their participation on an hourly basis and
477 received a bonus according to their performance (proportional to the sum of all the collected
478 apples' size). One subject from the dopamine group was excluded due to not engaging in the task
479 and performing at chance level. The study was approved by the UCL research ethics committee
480 and all subjects provided written informed consent.

481 *Pharmacological manipulation*

482 To reduce noradrenaline functioning, we administered 40mg of the non-selective β-
483 adrenoceptor antagonist propranolol 60 minutes before test (Fig 1D). To reduce dopamine
484 functioning, we administered 400mg of the selective D2/D3 antagonist amisulpride 90 minutes
485 before test. Because of different pharmacokinetic properties, drugs were administered at different

times. The placebo group received placebo at both time points, in line with our previous studies (*41*, *46*, *65*).

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). Choice options were displayed until a key was pressed. 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. The subjects were instructed to make the most juice and that they would receive cash bonus according to their performance.

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 (*32*, *53*, *67*, *68*), 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 groups (Figure 1c) characterised by different means μ_i and number of initial samples. The rewards (apple sizes) for each bandit were sampled from a distribution, meaning that values were chosen with likelihood defined by a probability density function. For each bandit, this was 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). The 'certain standard bandit' provided three initial samples and 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. 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 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 be able to measure choice consistency. 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 make sure that subjects understood that the apples from the same tree were always of similar size (generated following a normal distribution), they played 10 training trials. In this training, based on three

532 displayed apples of similar size, they had to guess, between two options, which apple was the most
533 likely to come from the same tree and received feedback from their choice.

534 *Statistical analyses*

535 To ensure consistent performance across all subjects, we excluded one outlier subject
536 (belonging to the dopamine group) from our analysis due to not engaging in the task and
537 performing at chance level (defined as randomly sampling from one out of three bandits randomly,
538 i.e. 33%). Each bandits' selection frequency for a horizon condition were computed over all 200
539 trials and not only over the trials were this specific bandit was present (i.e. 3/4 of 200 = 150 trials).
540 We compared behavioural measures and model parameters using (paired-samples) t-tests and
541 repeated-measures (rm-) ANOVAs with the between-subject factor drug group (noradrenaline
542 group, dopamine group, placebo group) and the within-subject factor horizon (long, short).
543 Information seeking, expected values and scores were analysed using rm-ANOVAS with the
544 within-subject factor horizon. Measures that were horizon-independent (e.g. prior mean), were
545 analysed using one-way ANOVAs with the between-subject factor drug group. As drug groups
546 differed in negative affect (cf. Appendix 2 Table 1), which, through its relationship to anxiety (*69*),
547 is thought to affect cognition (*70*) and potentially exploration (*71*), we corrected for negative affect
548 (PANAS) in addition to IQ (WASI) in all ANOVAS by adding those two measures as covariates
549 (cf. Appendix 1 for analysis without covariates). We report effect sizes using partial eta squared
550 (η2) for ANOVAs and Cohen's d (d) for t-tests (*72*).

551 *Computational modelling*

552 We adapted a set of Bayesian generative models from previous studies (*1*), where each
553 model assumed that different characteristics account for subjects' behaviour. The binary indicators
554 (c_{tr}, c_n) indicate which components (tabula-rasa 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 bandit 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 (*73*) 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 the so-called random exploration). The Thompson model reflects Thompson sampling (captures an uncertainty-dependent 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 tabula-rasa 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 5a). 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 2) 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:

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

Where the coefficient c_n adds the novelty exploration component. The probability of choosing bandit i was given by passing this into the softmax decision function:

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

where β is the inverse temperature of the softmax (lower values producing more stochasticity), and the coefficient c_{tr} adds the tabula rasa 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 (*74*)). 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:

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

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

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:

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

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 2) 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 tabula-rasa 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 was used (equivalent to performing MLE). A summary of the parameter values per group and per horizon can be found in Appendix 2 Table 2.

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

634 selection, we used maximum likelihood estimation (MLE), where we maximised the likelihood
635 for each subject individually (fmincon was ran with 8 randomly chosen starting point to overcome
636 potential local minima). We fitted the model using K-1 folds and validated the model on the
637 remaining fold. We repeated this process K times, so that each of the K fold is used as a validation
638 set once, and averaged the likelihood over held out trials. We did this for each model and each
639 subject and averaged across subjects. The model with the highest likelihood of held-out data (the
640 winning model) was the Thompson sampling with $(\mathrm{c_{tr}}, \mathrm{c_n}) = \{1,1\}$. It was also the model which
641 accounted best for the largest number of participants (Figure 4 – Figure supplement 1).

642 *Parameter recovery.*

643 To make sure that the parameters are interpretable, we performed a parameter recovery analysis.
644 For each parameter, we took 4 values, equally spread, within a reasonable parameter range ($\sigma_0 =$
645 $[0.5, 2.5], Q_0 = [1, 6], \epsilon = [0, 0.5], \eta = [0, 5]$). All parameters but Q_0 were free to vary as a
646 function of the horizon. We simulated behaviour on new trial sets using every combination (4^7),
647 fitted the model using MAP estimation (cf. Parameter estimation) and analysed how well the
648 generative parameters (generating parameters in Figure 5) correlated with the recovered ones
649 (fitted parameters in Figure 5) using Pearson correlation (summarised in Figure 5c). In addition to
650 the correlation we examined the spread (Figure 4 – Figure supplement 3) of the recovered
651 parameters. Overall the parameters were well recoverable.

652 *Model validation*

653 To validate our model, we used each participant's fitted parameters to simulate behaviour on our
654 task (4000 trials per agent). The stimulated data (Figure 5 – Figure supplement 1), although not
655 perfect, resembles the real data reasonably well. Additionally, to validate the behavioural
656 indicators of the two different exploration heuristics we stimulated the behaviour of 200 agents

each with the winning model. For the indicators of tabula-rasa 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 tabula-rasa 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 ($\sigma_0 = 1.312, \epsilon = 0.1, Q_0 = 3.2$).

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, and a 2018 NARSAD Young Investigator Grant (27023) from the Brain and Behavior Research Foundation. 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, in *Journal of Machine Learning Research* (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. R. S. Sutton, A. G. Barto, Introduction to Reinforcement Learning. *MIT Press Cambridge* (1998), doi:10.1.1.32.7692.
12. 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).
13. 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).
14. M. Silvetti, E. Vassena, E. Abrahamse, T. Verguts, *Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner* (2018), vol. 14.
15. 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).
16. A. J. Yu, P. Dayan, Uncertainty, neuromodulation, and attention. *Neuron*. **46**, 681–692 (2005).
17. 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.
18. J. David Johnson, Noradrenergic control of cognition: global attenuation and an interrupt function. *Med.*

Hypotheses (2003), doi:10.1016/s0306-9877(03)00021-5.

19. 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.
20. P. Dayan, A. J. Yu, Phasic norepinephrine: A neural interrupt signal for unexpected events. *Netw. Comput. Neural Syst.* (2006), doi:10.1080/09548980601004024.
21. 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* (2014), doi:10.1016/j.cell.2014.08.037.
22. 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.
23. 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.
24. S. J. Gershman, Y. Niv, Novelty and Inductive Generalization in Human Reinforcement Learning. *Top. Cogn. Sci.* **7**, 391–415 (2015).
25. 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).
26. 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.
27. 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).
28. 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.
29. 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.
30. E. Düzel, W. D. Penny, N. Burgess, Brain oscillations and memory. *Curr. Opin. Neurobiol.* (2010), , doi:10.1016/j.conb.2010.01.004.
31. 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.
32. 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*. **12**, 1–18 (2017).
33. D. Wechsler, WASI -II: Wechsler abbreviated scale of intelligence - second edition. *J. Psychoeduc. Assess.* (2013), doi:10.1177/0734282912467756.
34. S. Agrawal, N. Goyal, Analysis of thompson sampling for the multi-armed bandit problem. *J. Mach. Learn. Res.* **23**, 1–26 (2012).
35. 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.
36. 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).
37. G. Aston-Jones, J. D. Cohen, AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance. *Annu. Rev. Neurosci.* (2005), doi:10.1146/annurev.neuro.28.061604.135709.
38. D. Servan-Schreiber, H. Printz, J. D. Cohen, A network model of catecholamiine effects: Gain, signal-to-noise ratio, and behavior. *Science (80-.).* (1990), doi:10.1126/science.2392679.
39. 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).
40. 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.
41. 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.
42. 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.
43. 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.
44. 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).
45. 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).
46. 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.
47. Z. L. Rossetti, S. Carboni, Noradrenaline and dopamine elevations in the rat prefrontal cortex in spatial working memory. *J. Neurosci.* **25**, 2322–2329 (2005).
48. 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).
49. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science (80-.).* **275**, 1593–1599 (1997).
50. B. C. Wittmann, N. D. Daw, B. Seymour, R. J. Dolan, Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron*. **58**, 967–973 (2008).
51. 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).
52. 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).
53. W. K. Zajkowski, M. Kossut, R. C. Wilson, A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife* (2017), doi:10.7554/elife.27430.
54. 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).
55. 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).
56. 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.
57. 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.
58. 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.
59. T. Kahnt, P. N. Tobler, Dopamine Modulates the Functional Organization of the Orbitofrontal Cortex. *J. Neurosci.* (2017), doi:10.1523/jneurosci.2827-16.2016.
60. 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.
61. 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.
62. 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.
63. 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).
64. 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* (2014), doi:10.1001/jamapsychiatry.2014.1093.
65. 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.
66. 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.
67. D. Guo, A. J. Yu, in *Advances in Neural Information Processing Systems* (2018).
68. 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).
69. 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).
70. S. J. Bishop, C. Gagne, Anxiety, Depression, and Decision Making: A Computational Perspective. *Annu. Rev. Neurosci.* **41**, 371–388 (2018).
71. 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).
72. 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.
73. C. M. Bishop, in *Information Science and Statistics* (2006).
74. M. Speekenbrink, E. Konstantinidis, Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* (2015), doi:10.1111/tops.12145.

Appendix 1

Reaction times

There were no differences in reaction times (RT) between drug groups. Neither in the average RT (ANOVA: F(2, 54)=1.625, p=.206, η2=.057) nor in the RT in each horizon condition (rm-ANOVA: 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).

Dopamine effect on high-value bandit sampling frequency

The dopamine group had a marginal tendency towards selecting the high-value bandit, meaning that they were disposed to exploit more overall (noradrenaline 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 participants 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 participants have a larger σ_0 compared to females in the dopamine group, (t(19)=-2.466, p=.025, d=1.124; noradrenaline 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 reaction times (RT), driven primarily by female participants tending to have a faster RT in the long horizon compared to male participants (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 that 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 tabula-rasa 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 tabula-rasa 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), and in the consistency (F(2, 56)=3.596, p=.034, η2=.114; Pairwise comparisons: placebo vs noradrenaline: t(40)=-3.525, p=.001, d=.788; dopamine vs placebo: t(38)=1.107, p=.272, d=.251; dopamine vs noradrenaline: t(38)=-2.267, p=.026, d=.514), as well as in the model parameter for tabula-rasa exploration (ϵ: F(2, 56)=3.205, p=.048, η2=.103; Pairwise comparisons: placebo vs noradrenaline: t(40)=3.177, p=.002, d=.71; dopamine vs placebo: t(38)=.251, p=.802, d=.057; dopamine vs noradrenaline: t(38)=2.723, p=.009, d=.626).

Appendix 2

	Noradrenaline	Placebo	Dopamine	
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 noradrenaline: t(56)=2.801, p=.007, d=.799; dopamine vs placebo: t(56)=-2.096, p=.041, d=.557; dopamine vs noradrenaline: t(56)=.669, p=.506, d=.383). This was accounted for in the data analysis. Mean (SD).

Model	**Thompson**				**UCB**				**Hybrid**			
		$+\epsilon$	$+\eta$	$+\epsilon+\eta$		$+\epsilon$	$+\eta$	$+\epsilon+\eta$		$+\epsilon$	$+\eta$	$+\epsilon+\eta$
Fixed	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 depen-dent	σ_0	σ_0, ϵ	σ_0, η	σ_0, ϵ, η	γ, β	γ, β, ϵ	γ, β, η	$\gamma, \beta, \epsilon, \eta$	σ_0, γ, β	$\sigma_0, \gamma, \beta, \epsilon$	$\sigma_0, \gamma, \beta, \eta$	$\sigma_0, \gamma, \beta, \epsilon, \eta$
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												

Appendix 2 Table 2.

Table of parameters used for each model compared during model selection. 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.

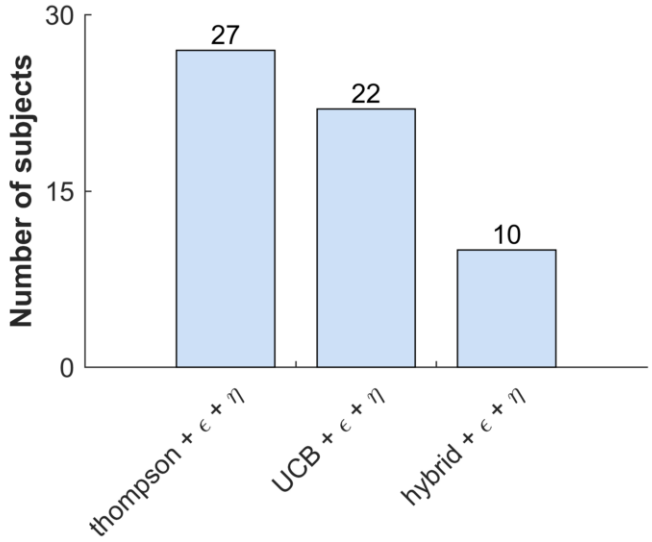

Figure 4 - Figure supplement 1

Model comparison: further evaluations. 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 participants.

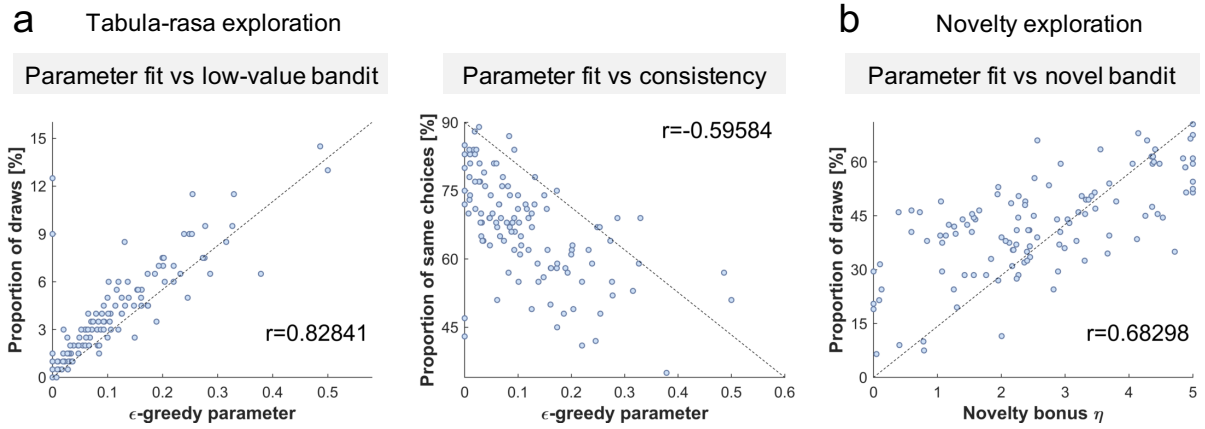

Figure 4 - Figure supplement 2

Correlations between model parameters and behaviour. The behavioural indicators of (a) tabula-rasa 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.

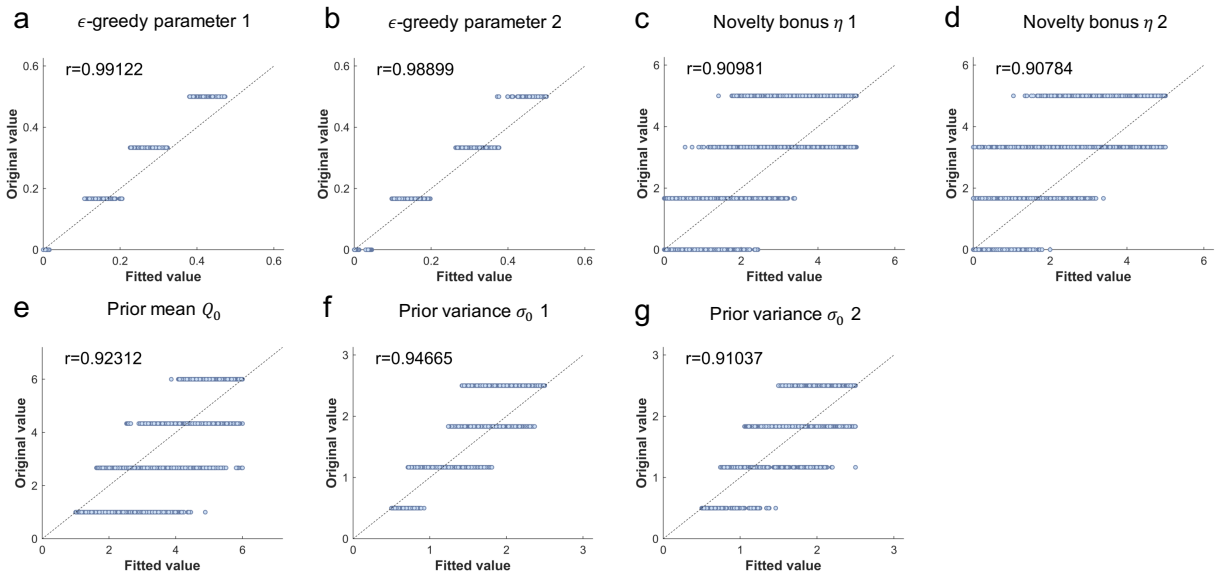

Figure 4 - Figure supplement 3

Parameter recovery analysis details. For each parameter of the winning model, we took 10 values, equally spread within the parameter range. We simulated behaviour using every combination, fitted the model and analysed how well the generative parameters (original values) correlated with the recovered ones (fitted parameters). Pearson correlation coefficient =r.

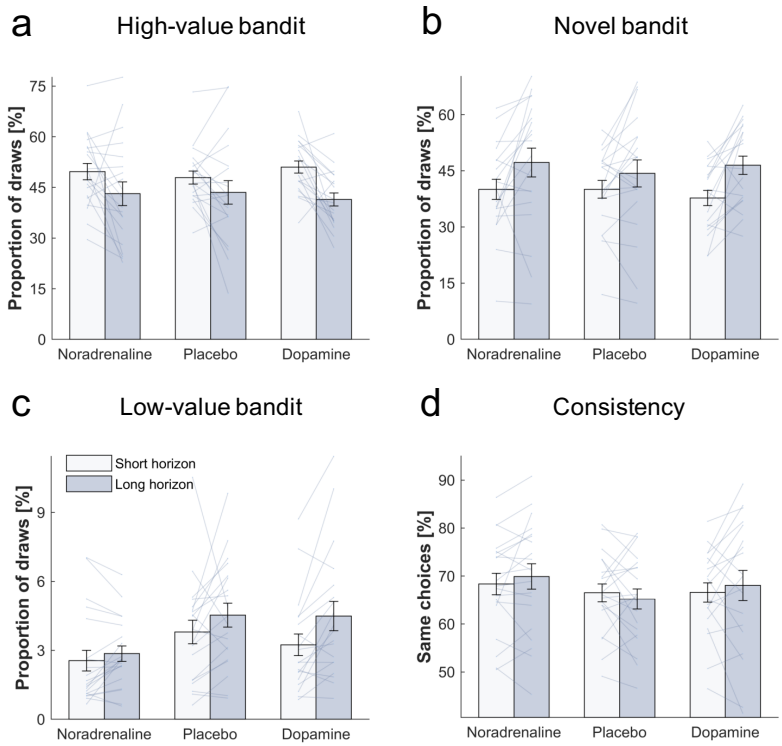

Figure 5 - Figure supplement 1

Simulated behaviour. We used each participant's 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.

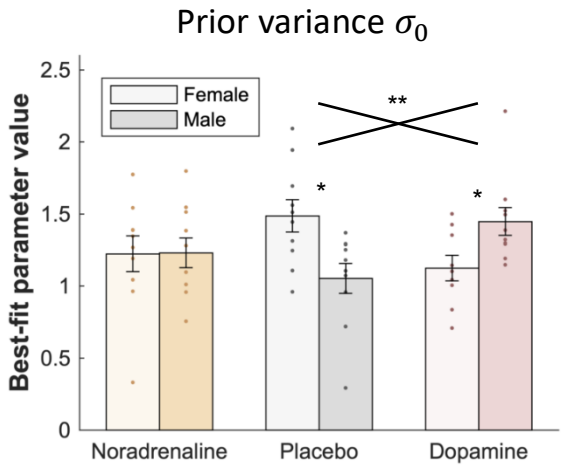

Figure 5 – Figure supplement 2

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