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: 30
Number of Figures: 5
Number of Tables: 0
Abstract: 148 words
Introduction: 644 words
Discussion: 750 words

Conflict of interest: The authors declare no competing financial interests.

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).

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 complex and computationally demanding exploration algorithms. Given known limitations in human cognitive resources, we hypothesised that humans would deploy additional, heuristic exploration strategies that carry the benefit of being computationally efficient. We examine the contributions of such heuristics in human choice behaviour and using computational modelling show this involves a tabula-rasa exploration that ignores all prior knowledge and a novelty exploration that targets completely unknown choice options respectively. In a double-blind, placebo-controlled pharmacological study, assessing contributions to the task of modulatory influences from catecholamines 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 know you like (chocolate), or one you have never tried? 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 lesser known options (exploration) so as to not miss out on possible higher rewards. In the latter case, you can opt to either chose an option that resembles your favourite (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 practice for studying complex decision making problems, such as 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. Such approaches reveal humans use strategies that reflect implementation of computationally demanding exploration algorithms[1,2]. One such strategy is to award an 'information bonus' to choice options, where this is scaled to their uncertainty. This is captured in algorithms like Upper Confidence Bound[3,4] (UCB) and leads to directed exploration of choice options agents knowns little about[1,5] (e.g. the hibiscus ice-cream). An alternative strategy, as captured by Thompson sampling[6] (and sometimes referred to as random exploration[7]), is to induce stochasticity that reflects uncertainty about choice options, which increases the likelihood of choosing options slightly inferior to the one you believe is best (e.g. the Toblerone ice-cream).

Both the above processes are computationally demanding, especially in real-life multiple-alternative decision problems, and may exceed one's available cognitive resources[8–10]. Indeed, the computational algorithms that have been suggested (e.g. UCB and Thompson

sampling) require computing and representing the expected value of all possible choice options plus their respective uncertainties (or variance). Here, we propose and examine the explanatory power of two additional computationally less costly forms of exploration, namely tabula-rasa and novelty exploration.

Computationally the most efficient way to explore is to ignore all prior information and to choose entirely randomly, de facto assigning the same probability to all options. Such 'tabula-rasa exploration' forgoes costly computation and is a reflection of what is known in reinforcement learning as an ε-greedy algorithmic strategy[11]. The computational efficiency, however, comes at a 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, tabula-rasa exploration has neurobiological plausibility, and the neurotransmitter noradrenaline has been ascribed the function of a 'reset button' that interrupts ongoing information processing[12–14]. Prior experimental work in rats shows that boosting noradrenaline leads to more tabula-rasa-like behaviour[15]. Here, we examined 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. Human studies support the presence of such novelty seeking[16–19]. This strategy can be implemented by a low-cost version of a UCB algorithm (i.e. does not rely on precise uncertainty estimates), in which a novelty bonus[20] is only added to previously unseen choice options. The neurotransmitter dopamine is implicated in signalling this type of novelty bonus, where evidence indicates its plays a role in processing and exploring novel and salient states[17,21–24].

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, where 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 (Fig. 1; bandits depicted as trees), enabling us to assess the contributions of tabula-rasa and novelty exploration in addition to Thompson sampling and UCB. 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 prior draws) or no information (0 prior draws). Novelty exploration predicts that a novel option will be chosen disproportionally more often (Fig. 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 (Fig. 1e). A second prediction is that the choice consistency is substantially affected by tabula-rasa exploration. Given that tabula-rasa exploration ignores all prior information, it is less likely to choose again the same bandit even when facing identical choice options (Fig. 1e). This contrast to other choice strategies that make consistent exploration predictions (e.g. UCB would consistently explore the choice options that carry a high information bonus).

We generated bandits from four different generative groups (Fig. 1c) with distinct sample means (fixed sampling variance) and prior reward information (i.e., number of previous draws). 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 prior samples (depicted as apples; similar to the horizon task[7]). The 'low-value bandit' was a bandit with one 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 sample was shown, appealing to a novelty exploration strategy. To assess choice consistency, all trials were repeated once. To avoid some exploration strategies overshadowing 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. Thus, subjects could perform either one draw encouraging exploitation (short horizon condition) or six draws encouraging more substantial exploration behaviour (long horizon condition)[7,25],

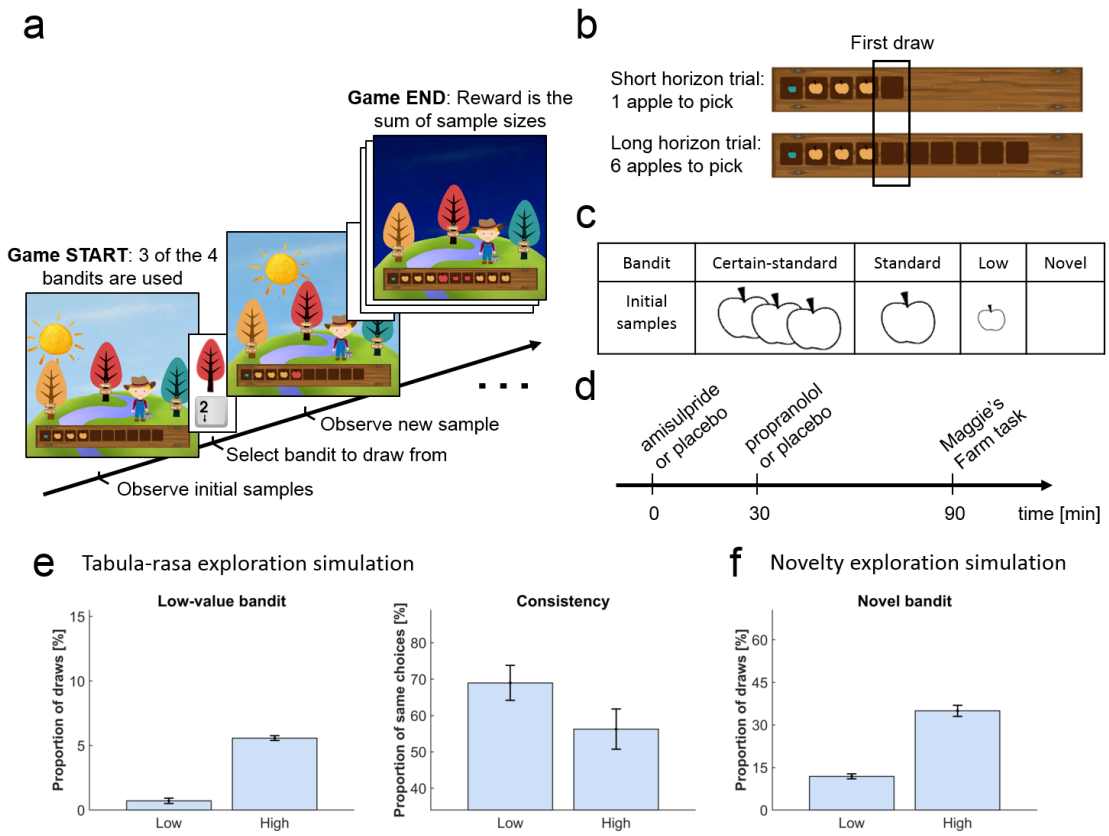

Figure 1: Study design. In the Maggie's farm task, subjects had to choose from three bandits (depicted as trees) to maximise an outcome (apple size). The samples (apples) of each bandit followed a normal distribution with a fixed sampling variance. (a) At the beginning of each trial, subjects were

provided with initial bandit samples on the wooden crate at the bottom of the screen, and had to select which bandit they want to sample from next. (b) Depending on the condition, they could either perform one draw (short horizon) or six draws (long horizon). The empty spaces on the wooden crate (and the sun 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 the number of samples initially shown. The 'certain-standard bandit' and the 'standard bandit' had comparable means but different levels of uncertainty in 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 in the dopamine condition, 40mg propranolol in the noradrenaline condition, or inert substances in the placebo condition. 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 were 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 (Fig. 1d) and compared to a placebo group that received a placebo at both drug times to match the corresponding antagonist time. One subject (amisulpride group) was excluded from analysis due to a lack of engagement with the task (cf. methods for details).

Increased exploration when information can be exploited

Our task embodied two potential time horizons, 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 (irrespective of drug condition). A marker of exploration here is evident if subjects chose a bandit with a lower

expected value (computed as the mean value of its initial samples shown; looking at the first draw, in accordance with the horizon task[7]). We found that subjects chose bandits with a lower expected value (i.e. they exploited less) in a long compared to a short horizon (repeated-measures ANOVA horizon main effect: $F(1,56)=18.988$, $p<.001$; drug main effect: $F(2,56)=1.361$, $p=.265$; interaction: $F(2,56)=.379$, $p=.686$; Fig. 2a). When we analysed the proportion of how often high-value option was chosen (i.e. the bandit with the highest expected reward based on the initial samples) we found subjects sampled from the high-value bandit more in the short compared to the long horizon (horizon main effect: $F(1,56) = 44.844$, $p<.001$; drug main effect: $F(2,56) = .181$, $p=.835$; interaction: $F(2,56)=2.494$, $p=.092$; Fig. 3a), confirming a reduction in exploitation when this information could be used subsequently. This behaviour resulted in a lower reward on the 1st sample in the long compared to the short horizon (horizon main effect: $F(1,56)=23.302$, $p<.001$; drug main effect: $F(2,56)=.487$, $p=.617$; interaction: $F(2,56)=1.551$, $p=.221$; Fig. 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 (horizon main effect: $F(1,56)=47.227$, $p<.001$; drug main effect: $F(2,56)=.206$, $p=.815$; interaction: $F(2,56)=1.207$, $p=.307$; Fig. 2b).

Next to evaluate whether subjects used additional information in the long horizon condition beneficially, we compared average reward (across six draws) obtained in the long compared to short horizon (one draw). We found average reward was higher in the long horizon (horizon main effect: $F(1,56)=105.165$, $p<.001$; drug main effect: $F(2,56)=.286$, $p=.752$; interaction: $F(2,56)=.678$, $p=.512$; Fig. 2c), indicating subjects tended to choose less optimal bandits at first but subsequently learnt to appropriately exploit harvested information to guide a choice of better bandits in the long run. Neither behaviour differed between drug groups, indicating neither noradrenaline nor dopamine influenced the degree of exploitation.

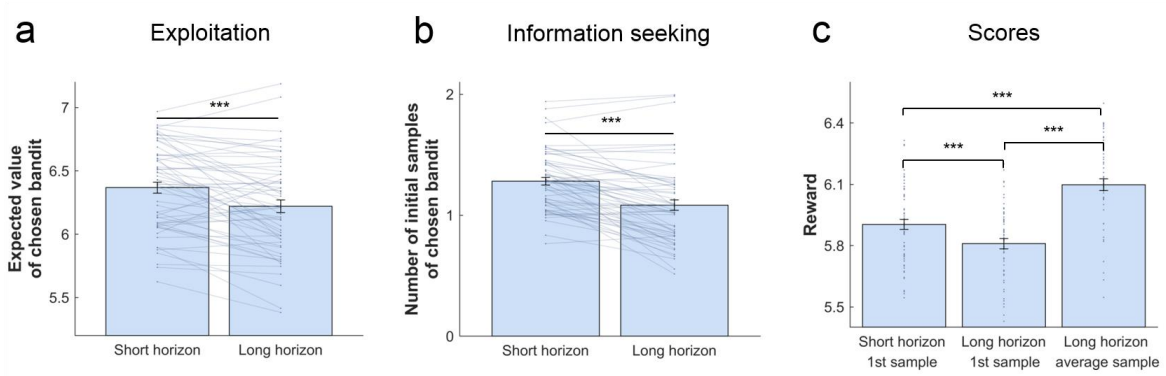

Figure 2: Benefits of exploration. To investigate the effect of information on performance we collapsed subjects over all three treatment groups. (a) The first bandit subjects chose as a function of its expected value (average of the samples that were initially revealed from it). Subjects chose bandits with a lower expected value (i.e. they exploit 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. This means that subjects sacrificed larger initial outcomes for the benefit of more information. This additional information helped to make better decisions in the long run, so that subjects earned more over all draws in the long horizon. *** = p<.001. Data are shown as mean ± SEM and each dot/line represents a subject.

Subjects show tabula-rasa behaviour

Tabula-rasa exploration (analogue to ϵ-greedy) predicts that ϵ % of the time each option will have equal probability of being chosen, a strategy that predicts tabula-rasa exploration will increase how often bad bandits (here: low-value bandit, cf. Fig 1e) are chosen. We investigated whether such behavioural signatures were increased in the long horizon condition when exploration is useful we found a highly significant main effect of horizon ($F(1,56)=24.240$, $p<.001$; Fig. 3c). This demonstrates this exploration strategy is utilised more when exploration is beneficial in the long horizon.

Tabula-rasa behaviour is modulated by noradrenaline function

When we tested whether tabula-rasa exploration was sensitive to neuromodulatory influences, we found the noradrenaline group chose the low-value option significantly less often than the placebo group, both in the short (placebo vs noradrenaline $t(38)=.389$, $p=.048$; Fig. 3c) and long horizon conditions (placebo vs noradrenaline $t(38)=2.102$, $p=.043$; Fig. 3c).

This was reflected also as an overall drug main effect (rm-ANOVA comparing noradrenaline and placebo group: drug main effect: $F(1,38) = 4.986$, $p=.032$; horizon main effect: $F(1,38)=9.614$, $p=0.004$; drug-by-horizon interaction: $F(1,38) = .037$, $p=.849$; when comparing all three groups: $F(2,56) = 2.523$, $p=.089$; interaction: $F(2,56)=1.744$, $p=.184$; Fig. 3c). These findings show that a key feature of tabula-rasa exploration, the choice of low-value bandits, is sensitive to influences from noradrenaline, with propranolol likely attenuating this influence.

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, it should lead to decreased choice consistency in relation to presentation of identical choice options. Because we repeated each trial twice in each horizon condition, we could compute consistency as the percentage of time subjects sampled from an identical bandit when facing the exact same choice options. As in the above analysis, we found the noradrenaline group chose significantly more consistently than the placebo group in both the short horizon (placebo vs noradrenaline $t(38)=-2.720$, $p=.034$) and the long horizon (placebo vs noradrenaline $t(38)=-2.195$, $p=.010$). This was also reflected by a drug main effect (rm-ANOVA comparing noradrenaline and placebo group: drug main effect: $F(1,38) =7.429$, $p=.010$; horizon main effect: $F(1,38)=.160$ $p=.691$; drug-by-horizon interaction: $F(1,38) = 1.700$, $p=.200$; when comparing all three groups: drug main effect: $F(2,56)=3.596$, $p=.034$; horizon main effect: $F(1,56)=.716$, $p=.401$; horizon-by-drug interaction: $F(2.56)=2.823$, $p=.068$; Fig. 3d). Taken together, these results indicate that tabula-rasa exploration depends critically on noradrenaline function, such that an attenuation of noradrenaline leads to a reduction in tabula-rasa exploration.

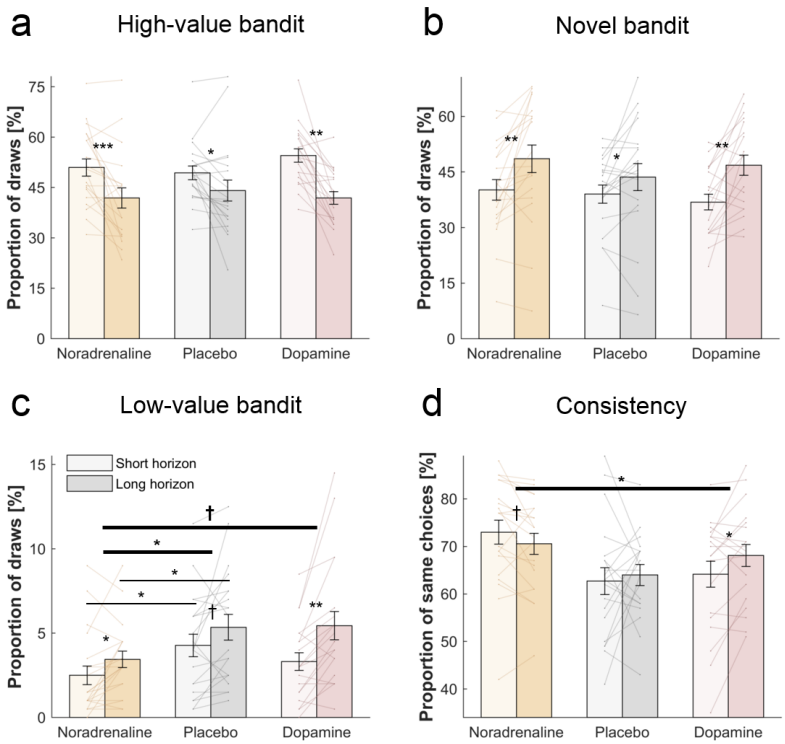

Figure 3: Behavioural horizon and drug effects. Choice patterns in the first draw for each horizon and drug group (noradrenaline, placebo and dopamine). Subjects from all three groups sampled from the (a) high-value bandit (i.e. bandit with the highest prior mean) more in the short horizon compared to 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 independent t-tests (thin). † = p<.10 (across all drug groups), * = p<.05, ** = p<.01, *** = p<.001. Data are shown as mean ± SEM and each dot/line represents one participant.

Novelty exploration unaffected by catecholaminergic drugs

Next, we examined whether subjects show evidence for novelty exploration by choosing novel bandit for which there was no prior information, as predicted by model simulations (Fig. 1g). We found a significant main effect of horizon (F(1,56)=30.867, p<.001; Fig. 3b) consistent with subjects exploring novel bandit significantly more often in a long horizon condition. Assessing whether novelty exploration was sensitive to drug manipulations, we found no drug effects on the novel bandit (drug main effect: F(2,56) =.352, p=.705; horizon-by-drug

interaction: $F(2,56)=1.331$, $p=.272$; Fig. 3b), indicating that an attenuation of dopamine or noradrenaline function does not impact novelty exploration in this task.

Subjects combine computationally demanding strategies with tabula-rasa and novelty exploration

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 directed (UCB) exploration, Thompson sampling, tabula-rasa (ε-greedy) exploration, novelty (novelty bonus) exploration, and combinations thereof (cf. Methods). Essentially, each model makes different exploration predictions. In computationally demanding Thompson sampling[6,26], both expected value and uncertainty contribute to choice, where higher uncertainty leads to more exploration. Instead of selecting the bandit with the highest mean, bandits are chosen relative to how often a random sample would yield the highest outcome, thus accounting for uncertainty in the bandits[2]. In UCB exploration[3,4], each bandit is chosen according to a mixture of expected value and the 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 option's value, the higher the information gain). The novelty exploration is a simplified version of the information bonus in UCB, which only applies to entirely novel options. It defined an intrinsic source of value in 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 in ε % of the time, irrespective of the prior information of this bandit.

We used cross-validation for model selection (Fig. 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%). Interestingly, as in previous studies[1], the hybrid model combining UCB and Thompson sampling explained the data better as compared to UCB and Thompson sampling alone, but this was no longer the case when accounting for novelty and tabula-rasa exploration (Fig. 4a). The winning model further revealed that all parameter estimates could be accurately recovered (Fig. 4b). 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).

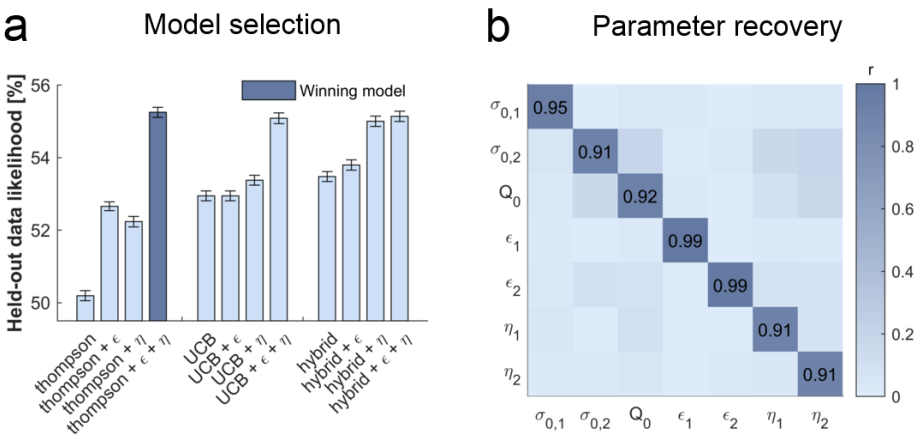

Figure 4: Subjects use a mixture of exploration strategies. (a) A 10-fold cross-validation of the likelihood of held-out data was used for model selection (chance level = 33%). The Thompson model with both the tabula-rasa ϵ-greedy parameter and the novelty bonus η best predicts held-out data. (b) Model simulation with 4^7 simulations predicted good recoverability of model parameters, σ_0 is the prior variance and Q_0 is the prior mean, both affecting Thompson sampling model; 1 stands for short, 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 the drug groups (Fig. 5, Table S2). First, we examined the ϵ-greedy parameter that captures the contribution of tabula-rasa exploration to choice behaviour. In accordance with the model-free analyses, we found this ϵ-greedy parameter to be higher in the long compared to the short horizon, consistent with an increased deployment of tabula-rasa exploration in the long horizon (horizon main effect: F(1,56)=10.362, p=.002).

Next, we assessed how this tabula-rasa exploration differed between drug groups. A significant drug main effect (across all three groups: F(2,56)=3.205, p=.048; horizon-by-drug interaction: F(56,2)=2.455, p=.095; Fig. 5a) demonstrates the drug groups differ in how strongly they deploy this exploration strategy. Post-hoc pairwise comparisons showed that subjects with reduced noradrenaline function had the lowest values of ϵ, primarily in the long horizon condition (placebo vs noradrenaline t(38)=2.490, p=.019; noradrenaline vs dopamine t(37)=2.672, p=.014; placebo vs dopamine t(37)=-.741, p=.464; short horizon: placebo vs noradrenaline t(38)=1.981, p=.055; noradrenaline vs dopamine t(37)=1.162, p=.253; placebo vs dopamine t(37)=.543, p=.591). In accordance with the behavioural results (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 S1), this demonstrates that noradrenaline blockade reduces tabula-rasa exploration.

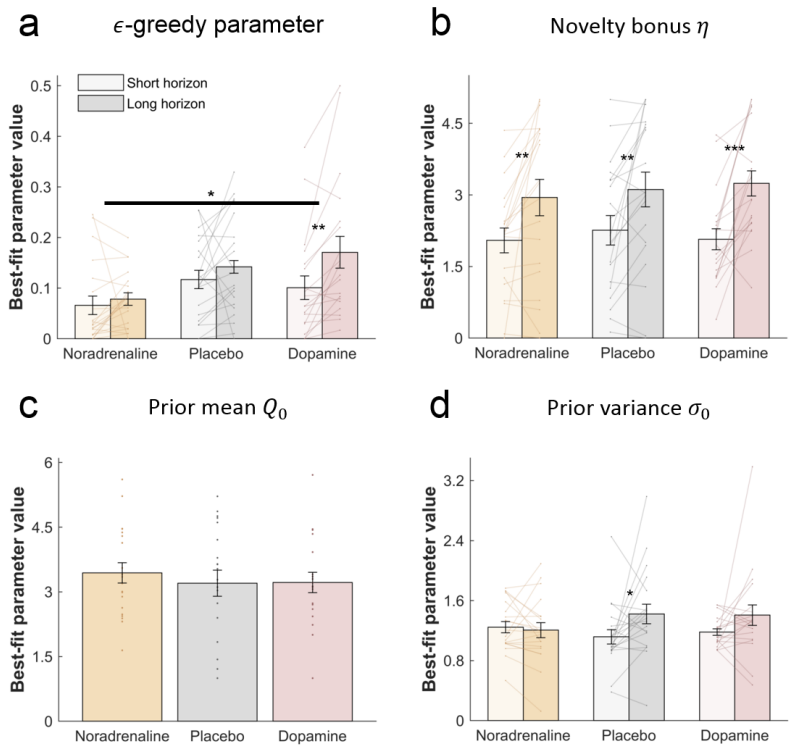

Figure 5: Drug effects on model parameters. The winning model's parameters were fitted to each subject's first draw. (a) All groups had higher values of ϵ (tabula-rasa exploration) in the long compared to the short horizon. Notably, subjects in the noradrenaline group had lower values of ϵ overall, indicating that attenuation of noradrenaline function reduces tabula-rasa exploration. Subjects from all groups (b) assigned a similar value to novelty 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 samples and (d) were similarly uncertain σ_0 about it although this uncertainty contributed more strongly in the long compared to the short horizon. * = p<.05, ** = p<.01, *** = p<.001.

No drug effects on other exploration forms

The novelty bonus η captures the intrinsic reward of selecting a novel option, over and above the contributions of uncertainty to Thompson (and UCB) sampling. When we examined whether the drug manipulation also affected other exploration strategies we found an increased novelty bonus effect in the long horizon (horizon main effect: F(1,56)=38.103, p<.001; Fig. 5b), showing that this exploration strategy was deployed more in a condition when exploration was of use. However, there was no difference between drug groups in terms of this effect (drug main effect: F(2,56)=.139, p=.871; horizon-by-drug interaction: F(2,56)=.379, p=.686), in line

with the model-free behavioural findings (correlation between the novelty bonus η with draws from the novel bandit: $R_{Pearson}$=.683, p<.001; Figure S1).

Lastly, we analysed parameters that governed the computationally demanding Thompson sampling. The initial estimate of a bandit's mean and its uncertainty are given by Q_0 (prior mean) and σ_0 (prior variance) respectively. The prior variance σ_0, which we allowed to vary between horizons, showed an effect of horizon (horizon main effect: F(1,56)=5.032, p=.029; Fig. 5d), demonstrating Thompson-based exploration was deployed more in the long horizon too. However, we did not find any drug effects for either parameter (prior variance σ_0: drug main effect: F(2,56)=.175, p=.840; horizon-by-drug interaction: F(2,56)=2.081, p=.134; prior mean Q_0; F(2,56)=.264, p=.769; Fig. 5c). Taken together, these findings show that tabula-rasa exploration was most sensitive to 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 expected reward (mean) of each choice option into account, a demand on computational resources which might not always be feasible[10]. Here, we demonstrate 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[20]) to an option not encountered before[27], a novelty bonus can be seen as an efficient simplification of demanding algorithms, such as UCB[3,4]. It is also interesting to note that the winning model did not include UCB, but novelty exploration. This indicates 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 both tabula-rasa and novelty exploration were deployed in a goal-directed manner, coupled with increased levels of exploration when this was strategically useful. Importantly, our results demonstrate that this tabula-rasa exploration is at work, even when accounting for what has previously been termed as 'random exploration'[7], which was captured by our more complex models.

It remains unclear how exploration strategies are implemented neurobiologically. Noradrenaline inputs, arising from locus coeruleus[28] (LC) are known to modulate exploration[2,29,30], though empirical data on its precise mechanisms and locus of action remains limited. In this study, we found noradrenaline impacted tabula-rasa exploration, suggesting

noradrenaline may disrupt ongoing valuation processes and discarding prior information. This is consistent with findings in rodents wherein enhanced anterior cingulate noradrenaline release led to more random behaviour[15]. It is also in line with pharmacological findings in monkeys that show enhanced choice consistency after reducing LC noradrenaline firing rates[31]. Our findings expand on previous human studies that showed an association between indirect markers of noradrenaline activity (i.e. pupil diameter[32]) are associated with exploration, but did not dissociate different potential exploration strategies subjects can deploy[33].

LC has two known modes of synaptic signalling[28]: tonic and phasic, and these are thought to have complementary roles[14]. Phasic noradrenaline is thought to act as a reset button[34], rendering an agent agnostic to all previously accumulated information, a de facto signature of tabula-rasa exploration. Tonic noradrenaline has been associated with increased exploration[29,35], decision noise in rats[36] and more specifically with random as opposed to directed exploration strategies[37]. 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, albeit to a lesser extent, affects dopamine function making it difficult to assign a precise function to noradrenaline. This latter influenced our decision to opt for drugs with high specificity for either dopamine or noradrenaline[38], thus enabling us to reveal highly specific effect on tabula-rasa exploration. Although the contributions of tonic and phasic noradrenaline signalling cannot be disentangled in our study, our findings do align with theoretical accounts and non-primate animal findings, indicating phasic noradrenaline promotes tabula-rasa exploration.

Dopamine has also been suggested to be important for arbitrating between exploration and exploitation[39,40], but findings are heterogeneous with reported dopaminergic effects on

random exploration[41], on directed exploration[22,42], or no effect at all[43]. We did not observe any effect of dopamine manipulation on any exploration strategy. Because previous studies found neurocognitive effects with the same dose[38,44–46], we think it is unlikely this just reflects an ineffective drug dose. One possible reason for our null finding is that our dopaminergic blockade targets D2/D3 receptors rather than D1 receptors, a limitation due to the fact that there are no specific D1 receptor blockers available in humans. An expectation of greater D1 involvement arises out of theoretical models[47] and a prefrontal hypothesis of exploration[42]. Upcoming, novel drugs[48] might be able help unravel a D1 contribution to different forms of exploration.

In conclusion, humans supplement computationally expensive exploration strategies with less resource demanding exploration heuristics, namely tabula-rasa and novelty exploration. Our finding that noradrenaline specifically influences tabula-rasa exploration demonstrates that distinct exploration strategies are under an influence from specific neuromodulators. The findings can enable a richer understanding of disorders of exploration, such as attention-deficit/hyperactivity disorder[49,50] and how aberrant catecholamine function leads to its core behavioural impairments.

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. Each subject was randomly allocated to one of three drug groups, controlling for an equal gender balance across all groups. 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. The groups consisted of 20 subjects each matched (Table S1) for age, mood (PANAS[51]) and intellectual abilities (WASI abbreviated version[52]). 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 dopamine 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 test (Fig 1d). To reduce dopamine functioning, we administered 400mg of the selective D2/D3 antagonist amisulpride 90 minutes 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[38,44,53].

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 trees that produced apples with different sizes in two different

horizon conditions (Fig. 1a-b). The sizes of apples they collected were summed and converted to an amount of juice (feedback) and subjects are informed about their performance at the end of every 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) and within trials the mean reward μ (i.e. apple size) and the uncertainty captured by the number of initial samples (i.e. number of apples shown at the beginning of the trial) of each option. Each bandit (i.e. tree) i was from one of four generative groups (Fig. 1c) characterised by different means μ_i and number of initial samples. The samples of each bandit were then taken from $N(\mu_i, 0.8)$, truncated to [2, 10], and rounded to the closest integer. 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) as well as the medium-

value option (i.e. the bandit with the second 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-tree combination making it a total of 100 different trials. They were then duplicated to be able to measure choice consistency, and each one was played these 200 trials each 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 trees' 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. Based on three apples produced by a single tree, they had to guess from two options, which apple was the most likely to come from the same tree and received feedback from their choice.

Statistical analyses

To ensure consistent performance across all subjects, we excluded one outlier subject (belonging to the dopamine 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 randomly, i.e. 33%). We compared behavioural measures and model parameters using (paired-samples) t-tests and repeated-measures ANOVAs with the between-subject factor drug group (noradrenaline group, dopamine group, placebo group) and the within-subject factor horizon (long, short).

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 (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 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 trees is presented, the expected mean Q and precision $\tau = \frac{1}{\sigma^2}$ of the corresponding tree 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[54] in stationary bandits.

We examined three base models: UCB, Thompson sampling and a hybrid of UCB and Thompson sampling. We computed three extensions of each of them by either adding tabula rasa exploration $(\mathrm{c_{tr}}, \mathrm{c_n}) = \{1,0\}$, or 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 Fig. 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 Table S2) are described in detail below.

Choice rules

UCB. In this model, an information bonus γ is added to the expected reward of each option, scaling with the option's uncertainty[1]. 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 choice rule:

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

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 sampling. In this model, 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[55]). 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 UCB and Thompson sampling (similar to Gershman, 2018). 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 UCB is used while if w=0 only Thompson sampling 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 (Table S2) as they capture different exploration forms: exploration directed towards information gain (information bonus γ in UCB), exploration directed towards novelty (novelty bonus η), exploration based on the injection of stochasticity (inverse temperature β in UCB and prior variance σ_0 in Thompson) 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 UCB or Thompson sampling 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 the supplements (Table S3).

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 participants (Fig. S2).

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 on new trial sets using every combination (4^7), fitted the model using MAP estimation (cf. Parameter estimation) and analysed how well the generative parameters (generating parameters in Fig. 5) correlated with the recovered ones (fitted parameters in Fig. 5) using Pearson correlation (summarised in Fig. 5c). In addition to the correlation we examined the spread (Fig. S3) of the recovered parameters. Overall the parameters were well recoverable.

Model validation

To validate our model, we used each participant's fitted parameters to simulate behaviour on our task (4000 trials per agent). The stimulated data (Fig. S4), although not perfect, resembles the real data reasonable well. Additionally to validate the behavioural 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 (ϵ=0) and high (ϵ=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 (Fig. 1f) and the choice consistency (Fig. 1e). Similarly, for the indicator of novelty exploration, we simulated behaviour with low (η=0) and high (η=2) values of the novelty bonus η to validate the use of the proportion of the novel-bandit selection (Fig. 1g). Again, the remaining parameters were set to the mean parameter fits ($\sigma_0 = 1.312, \epsilon = 0.1, Q_0 = 3.2$).

References

1. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition* **173,** 34–42 (2018).
2. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55,** 7–14 (2019).
3. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. in *Journal of Machine Learning Research* (2003). doi:10.1162/153244303321897663
4. Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R. & Auer, P. 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. Schwartenbeck, P. *et al.* Computational mechanisms of curiosity and goal-directed exploration. *Elife* (2019). doi:10.7554/eLife.41703
6. Thompson, W. R. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. **25,** 285–294 (1933).
7. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143,** 2074–2081 (2014).
8. Cohen, J. D., McClure, S. M. & Yu, A. J. 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).
9. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441,** 876–879 (2006).
10. Dezza, I. C., Cleeremans, A. & Alexander, W. 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. Sutton, R. S. & Barto, A. G. Introduction to Reinforcement Learning. *MIT Press Cambridge* (1998). doi:10.1.1.32.7692
12. David Johnson, J. Noradrenergic control of cognition: global attenuation and an interrupt function. *Med. Hypotheses* (2003). doi:10.1016/s0306-9877(03)00021-5
13. Bouret, S. & Sara, S. J. Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends Neurosci.* (2005). doi:10.1016/j.tins.2005.09.002
14. Dayan, P. & Yu, A. J. Phasic norepinephrine: A neural interrupt signal for unexpected events. *Netw. Comput. Neural Syst.* (2006). doi:10.1080/09548980601004024
15. Tervo, D. G. R. *et al.* Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell* (2014). doi:10.1016/j.cell.2014.08.037
16. Bunzeck, N., Doeller, C. F., Dolan, R. J. & Duzel, E. Contextual interaction between novelty and reward processing within the mesolimbic system. *Hum. Brain Mapp.* (2012). doi:10.1002/hbm.21288
17. Wittmann, B. C., Daw, N. D., Seymour, B. & Dolan, R. J. Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron* (2008). doi:10.1016/j.neuron.2008.04.027
18. Gershman, S. J. & Niv, Y. Novelty and Inductive Generalization in Human Reinforcement Learning. *Top. Cogn. Sci.* **7,** 391–415 (2015).
19. Stojic, H., Shulz, E., Analytis, P. P. & Speekenbrink, M. It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *PsyArXiv* (2018).
20. Krebs, R. M., Schott, B. H., Schütze, H. & Düzel, E. The novelty exploration bonus and its attentional modulation. *Neuropsychologia* (2009). doi:10.1016/j.neuropsychologia.2009.01.015
21. Bromberg-Martin, E. S., Matsumoto, M. & Hikosaka, O. Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. *Neuron* (2010). doi:10.1016/j.neuron.2010.11.022
22. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Dopamine modulates novelty seeking behavior during decision making. *Behav. Neurosci.* (2014). doi:10.1037/a0037128
23. Düzel, E., Penny, W. D. & Burgess, N. Brain oscillations and memory. *Current Opinion in Neurobiology* (2010). doi:10.1016/j.conb.2010.01.004
24. Iigaya, K. *et al.* The value of what's to come: neural mechanisms coupling prediction error and reward anticipation. *bioRxiv* (2019). doi:10.1101/588699
25. Warren, C. M. *et al.* The effect of atomoxetine on random and directed exploration in humans. *PLoS One* **12,** 1–18 (2017).
26. Agrawal, S. & Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. *J. Mach. Learn. Res.* **23,** 1–26 (2012).
27. Foley, N. C., Jangraw, D. C., Peck, C. & Gottlieb, J. Novelty enhances visual salience independently of reward in the parietal lobe. *J. Neurosci.* (2014). doi:10.1523/JNEUROSCI.4171-13.2014
28. Rajkowski, J., Kubiak, P. & Aston-Jones, G. Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Res. Bull.* **35,** 607–616 (1994).

29. Aston-Jones, G. & Cohen, J. D. 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
30. Servan-Schreiber, D., Printz, H. & Cohen, J. D. A network model of catecholamiine effects: Gain, signal-to-noise ratio, and behavior. *Science (80-.).* (1990). doi:10.1126/science.2392679
31. Jahn, C. I. *et al.* Dual contributions of noradrenaline to behavioural flexibility and motivation. 2687–2702 (2018).
32. Joshi, S., Li, Y., Kalwani, R. M. & Gold, J. I. 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
33. Jepma, M. & Nieuwenhuis, S. 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
34. Dayan, P. & Yu, A. J. Phasic norepinephrine: A neural interrupt signal for unexpected events. *Netw. Comput. Neural Syst.* **17,** 335–350 (2006).
35. Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J. & Aston-Jones, G. The role of locus coeruleus in the regulation of cognitive performance. *Science (80-.).* **283,** 549–554 (1999).
36. Kane, G. A. *et al.* Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cogn. Affect. Behav. Neurosci.* **17,** 1073–1083 (2017).
37. Warren, C. M. *et al.* The effect of atomoxetine on random and directed exploration in humans. *PLoS One* (2017). doi:10.1371/journal.pone.0176034
38. Hauser, T. U., Moutoussis, M., Purg, N., Dayan, P. & Dolan, R. J. Beta-Blocker Propranolol Modulates Decision Urgency During Sequential Information Gathering. *J. Neurosci.* (2018). doi:10.1523/jneurosci.0192-18.2018
39. Kayser, A. S., Mitchell, J. M., Weinstein, D. & Frank, M. J. Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology* **40,** 454–462 (2015).
40. Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife* (2017). doi:10.7554/elife.27430
41. Cinotti, F. *et al.* Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Sci. Rep.* **9,** 1–14 (2019).
42. Frank, M. J., Doll, B. B., Oas-Terpstra, J. & Moreno, F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat. Neurosci.* (2009). doi:10.1038/nn.2342
43. Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S. C. & Heekeren, H. R. 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
44. Hauser, T. U., Eldar, E., Purg, N., Moutoussis, M. & Dolan, R. J. Distinct roles of dopamine and noradrenaline in incidental memory. *J. Neurosci.* (2019). doi:10.1523/jneurosci.0401-19.2019
45. Kahnt, T., Weber, S. C., Haker, H., Robbins, T. W. & Tobler, P. N. Dopamine D2-Receptor Blockade Enhances Decoding of Prefrontal Signals in Humans. *J. Neurosci.* (2015). doi:10.1523/jneurosci.4182-14.2015
46. Kahnt, T. & Tobler, P. N. Dopamine Modulates the Functional Organization of the Orbitofrontal Cortex. *J. Neurosci.* (2017). doi:10.1523/jneurosci.2827-16.2016
47. Humphries, M. D., Khamassi, M. & Gurney, K. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci.* (2012). doi:10.3389/fnins.2012.00009
48. Soutschek, A. *et al.* Dopaminergic D1 Receptor Stimulation Affects Effort and Risk Preferences. *Biol. Psychiatry* (2019). doi:10.1016/j.biopsych.2019.09.002
49. Hauser, T. U., Fiore, V. G., Moutoussis, M. & Dolan, R. J. Computational Psychiatry of ADHD: Neural Gain Impairments across Marrian Levels of Analysis. *Trends Neurosci.* **39,** 63–73 (2016).
50. Hauser, T. U. *et al.* 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
51. Watson, D., Clark, L. A. & Tellegen, A. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *J. Pers. Soc. Psychol.* (1988). doi:10.1037/0022-3514.54.6.1063
52. Wechsler, D. WASI -II: Wechsler abbreviated scale of intelligence - second edition. *J. Psychoeduc. Assess.* (2013). doi:10.1177/0734282912467756
53. Hauser, T. U. *et al.* Noradrenaline blockade specifically enhances metacognitive performance. *Elife* (2017). doi:10.7554/eLife.24901
54. Bishop, C. M. Machine Learning and Pattern Recogniton. in *Information Science and Statistics* (2006).
55. Speekenbrink, M. & Konstantinidis, E. Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* (2015). doi:10.1111/tops.12145