Neuron

Article

Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain

Highlights

- Uncertainty-directed exploration considers the prospective benefit of new information.
- Novelty-directed exploration is myopic and motivated by inflated reward expectation.
- Option features are integrated by vmPFC to balance the explore/exploit trade-off.
- Integrating a mixture of strategies offers a tractable approximation of t optimal control.

Authors

Jeffrey Cockburn, Vincent Man,
William Cunningham,
John P. O'Doherty

Correspondence

jcockbur@caltech.edu

In brief

Cockburn et al. show that novelty and uncertainty are used by the human brain to guide distinct exploration strategies. Uncertainty-directed exploration considers the prospective benefit of new information, whereas novelty motivates exploration by inflating the brain's expectation of reward, offering a feasible decomposition of an otherwise intractable explore/exploit dilemma.

Cockburn et al., 2022, Neuron *110*, 1–12
August 17, 2022 © 2022 Elsevier Inc.
https://doi.org/10.1016/j.neuron.2022.05.025

CellPress

Neuron

CellPress

Article

Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain

Jeffrey Cockburn,[1,3,*] Vincent Man,[1] William Cunningham,[2] and John P. O'Doherty[1]

[1]Division of Humanities and Social Sciences, Caltech, Pasadena, CA, USA
[2]Department of Psychology, University of Toronto, Toronto, ON, Canada
[3]Lead contact
*Correspondence: jcockbur@caltech.edu

https://doi.org/10.1016/j.neuron.2022.05.025

SUMMARY

Both novelty and uncertainty are potent features guiding exploration; however, they are often experimentally conflated, and an understanding of how they interact to regulate the balance between exploration and exploitation has proved elusive. Using a task designed to decouple the influence of novelty and uncertainty, we identify separable mechanisms through which exploration is directed. We show that uncertainty-directed exploration is sensitive to the prospective benefit offered by new information, whereas novelty-directed exploration is maintained regardless of its potential advantage. Using a computational framework in conjunction with fMRI, we show that uncertainty-directed choice is rooted in an adaptive bias indexing the prospective utility of exploration. In contrast, novelty persistently promotes exploration by optimistically inflating reward expectations while simultaneously dampening uncertainty signals. Our results identify separable neural substrates charged with balancing the explore/exploit trade-off to foster a manageable decomposition of an otherwise intractable problem.

INTRODUCTION

Adaptive organisms are faced with a fundamental trade-off between choosing a familiar option that leads to a known reward or exploring less familiar alternatives in hopes of finding something better (Cohen et al., 2007). The explore/exploit dilemma presents an exceptional challenge with only a narrow range of circumstances in which an optimal solution is known (Gittins, 1979). Despite its importance to survival, very little is known about how the human brain tackles this problem or how those neural computations are manifested behaviorally. However, an emerging body of literature has highlighted the contributions of two variables driving exploration in the mammalian brain: visual novelty (Ennaceur and Delacour, 1988; Hughes, 2007; Daffner et al., 1998; Wittmann et al., 2008) and outcome uncertainty (Wilson et al., 2014; Badre et al., 2012; Gershman, 2018; Blanchard and Gershman, 2018; Trudel et al., 2020). Despite their significance, these variables are often conflated experimentally, and nothing is known about how they interact to regulate exploration. Here, we describe a bespoke behavioral task designed to distinguish these two variables. We empirically test a new computational framework that describes how these variables contribute to exploration by using both behavioral data and neural data measured with functional magnetic resonance imaging (fMRI).

The machine-learning literature offers a growing catalogue of practical approaches to the explore/exploit dilemma. One class of algorithm, commonly referred to as directed exploration, shepherds preference toward ‘‘informative’’ options. This notion has been formalized by the upper confidence bound (UCB) algorithm, where an optimistic attitude of the unknown is captured in the form of an uncertainty bias expressing the plausible benefit that might be encountered by exploring uncertain options (Agrawal, 1995; Katehakis and Robbins, 1995; Auer et al., 2002). An alternative strategy, commonly referred to as optimistic value initialization, boosts the initial reward expectation associated with novel options, compelling the learning agent toward new opportunities (Brafman and Tennenholtz, 2002; Ng et al., 1999). Motivational and environmental changes can be rapidly accommodated by algorithms that integrate various biases (e.g. an uncertainy bias) as part of the decision-making process, although this flexibility adds the additional computational costs of deriving the appropriate biases anew for each decision. In contrast, algorithms such as optimistic initialization fuse an exploratory motivation into the reward expectation to foster exploration through a singular cached value and thus offer a computationally efficient but less flexible strategy. Here, we aim to investigate the trade-off between flexibility and computational efficiency offered by these two strategies as an additional understudied dimension of human learning and decision making (Collins and Cockburn, 2020).

Investigations into human strategies of regulating the explore/exploit trade-off offer mixed results; uncertainty aversion

(Payzan-LeNestour et al., 2013), indifference (Daw et al., 2006), and uncertainty-seeking behavior (Trudel et al., 2020; Blanchard and Gershman, 2018) are reported among a range of individual differences (Badre et al., 2012; Frank et al., 2009). Notably, when expected value and uncertainty are decorrelated, uncertain options are preferentially sampled when the information they offer can be exploited in the future, suggesting a prospective quality to uncertainty-directed exploration (Wilson et al., 2014; Gershman, 2018). In contrast, several lines of evidence show that animals and humans alike exhibit a robust preference for novel options and sequences (Ennaceur and Delacour, 1988; Hughes, 2007; Daffner et al., 1998; Fantz, 1964; Kidd et al., 2012, 2014; Costa et al., 2019). A puzzling incongruity is the fact that although appetites for uncertainty vary widely, novelty robustly draws favor despite the fact that new options are themselves inherently uncertain.

Research probing the neural correlates of exploration offer additional details on how the human brain balances this trade-off. The frontopolar cortex (FPC) has been shown to track the relative uncertainty of the options being considered (Badre et al., 2012; Yoshida and Ishii, 2006) and the advantage of switching to an alternative course of action (Boorman et al., 2009; Bunge and Wendelken, 2009). Disrupting the FPC with transcranial magnetic stimulation (TMS) modulates uncertainty-directed choice (Beharelle et al., 2015), further implicating the role of the FPC in exploration. Additional studies have also highlighted the ventral medial prefrontal cortex (vmPFC), as moderating the switch between exploratory and exploitative dimentions of behavior (Blanchard and Gershman, 2018; Trudel et al., 2020; Domenech et al., 2020), suggestive of a multi-hub circuit concerned with balancing exploration and exploitation.

Studies investigating the neural underpinnings of novelty processing have shown that novel stimuli evoke phasic dopamine (DA) release (Horvitz et al., 1997), linking novelty processing to the reward prediction error learning (RPE) signal (Montague et al., 1996; Schultz, 1998). Novelty-induced phasic DA signals have been argued to reflect a shaping bonus that encourages exploration (Kakade and Dayan, 2002), and this shaping bonus has been funcitonally linked to learning and decision making by imaging results showing that RPE signals in ventral striatum reflect a bias consistent with optimistic initiation (Wittmann et al., 2008; Krebs et al., 2009).

Although exploration is believed to be multi-faceted, little is known about how these features coexist or, in particular, how the apparent tension between novelty and uncertainty is resolved. Complicating this issue is the fact that novelty and uncertainty are challenging to distinguish experimentally because novel options are maximally uncertain. We hypothesized that it would be possible to uncover the unique influence of these variables by decoupling their respective contributions to behavior as a function of task horizon. Specifically, we hypothesized that visual novelty operates as a simple stimulus-driven heuristic measure that guides behavior via optimistic initialization independently of the consequences of exploration. Conversely, uncertainty-directed exploration is thought to reflect prospective valuation and therefore should adapt to an approaching task horizon by downgrading the value of information as its potential benefit declines.

To test these ideas, we exposed human participants to a newly designed decision-making task in which, while undergoing fMRI, participants chose between options that varied in terms of novelty, uncertainty, and expected reward. Our task design offers two important experimental advances. First, it decoupled uncertainty and novelty by explicitly revaluing options, thereby varying uncertainty independently of visual novelty. Second, it offered a means of probing feature-specific adaptation by offering novel and uncertain options in varying proximity to the task horizon. We developed a computational framework to describe how novelty and uncertainty interact and guide exploration and leverage fMRI data to discriminate between different forms of the model in key regions of interest (ROIs) identified on the basis of their role in implementing exploration/exploitation computations; namely these ROIs were the ventral striatum, the ventromedial prefrontal cortex, and the frontal pole.

RESULTS

Participants (n = 32) performed a learning and decision-making task while undergoing fMRI. In brief, while undergoing four consecutive 15-min fMRI sessions (five learning contexts blocks per session), participants played 20 blocks of a finite horizon multi-armed bandit task designed to expose the influence of reward history, estimation uncertainty, and stimulus novelty on balancing the explore/exploit trade-off. On each trial, participants were asked to choose between two slot machines (see Figure 1A). Having made their choice, they were informed of the machines' payout of either $1US (win) or $0US (loss). Participants were instructed that each learning block, framed as a visit to a new casino, would consist of approximately 20 trials (between 18 and 23 trials) and that they should try to accumulate as much reward as they could. At the end of the experiment, one of the casinos they visited was chosen at random, and a performance bonus was calculated from this casino.

Each block offered a new learning context that included five visually unique slot machines, three of which were familiar stimuli that had been seen in previously visited casinos; the remaining two were novel stimuli that had not yet been shown (see Figure 1B). Participants were informed that each slot machine had a fixed probability of winning within a given casino, but values across casinos were independent. Trials were structured such that the two slot machines offered varied in terms of reward probability (expected value manipulation), the number of previous exposures (novelty manipulation), and the number of times they'd been sampled in the current block (uncertainty manipulation), allowing us to to systematically examine the influence of reward, novelty, and uncertainty across the task horizon (see Figure 1C). Importantly, computational model comparison shows that value learned in previous contexts did not influence behavior in the current block (exceedance probabilities = 0.03), and regression analyses demonstrate that within a block of trials there were no periods in which values learned in previous blocks influenced choice (all p values > 0.05, uncorrected for multiple comparisons; see STAR Methods "analysis of task comprehension"), demonstrating that participants were motivated and understood the task structure.

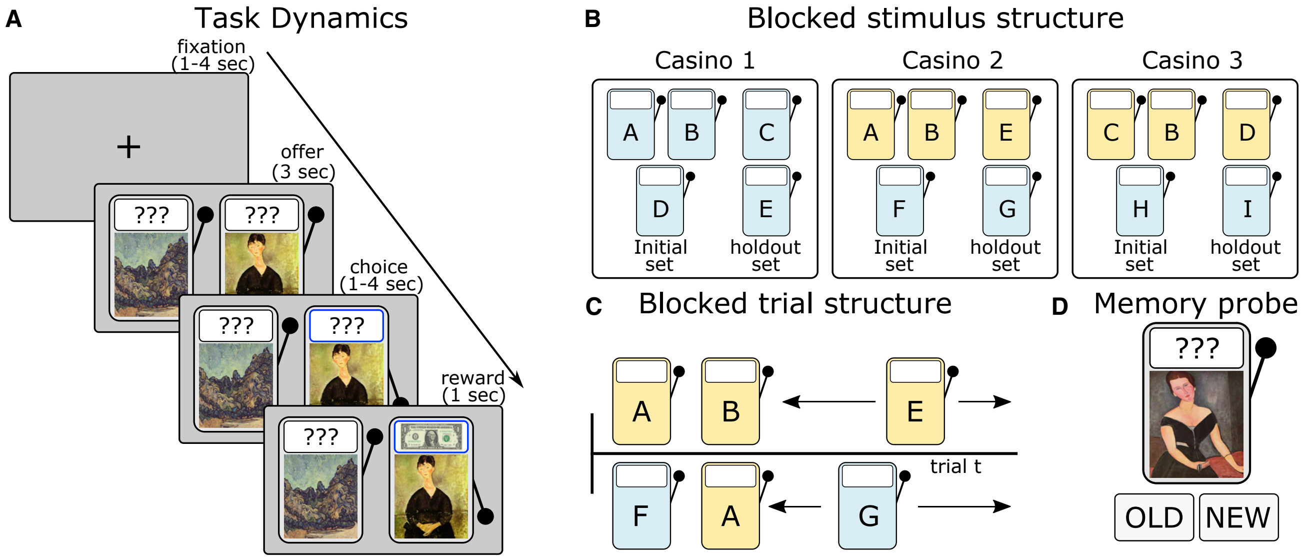

Figure 1. Experimental procedures

(A) Trial dynamics. After a jittered fixation screen (all jitters were sampled from a pseudo-randomized linear spacing of 1–4 s), participants were offered two slot machines and asked to select one by using a button box (left thumb to select the option on the left, right thumb to select the option on the right). Choice feedback was provided (handle moved to a pulled position, and outcome frame highlighted). After a 1–4 s jitter, reward feedback ($1/$0) was shown.

(B) Blocked stimulus structure: each learning context was framed as a visit to a unique casino. Five different slot machines could be offered in each casino, two of which were novel and three of which were familiar (blue and yellow, respectively, here for demonstration purposes—all machines using during the task were grey). Slot machines were segregated to form an initial set of three stimuli used early in each block, and a holdout set of two stimuli (one novel, the other familiar) that were gradually added to the initial set from which offered stimuli could be drawn on each trial.

(C) Blocked trial structure: offers were made with stimuli from the initial set at the start of each block. This included pairs with equal uncertainty but differing familiarity, as well as pairs with equal familiarity but differing uncertainty. At pseudo-random trials, the novel and familiar holdout stimuli were added to the set from which stimuli could be drawn.

(D) Memory probe task: having completed 20 blocks of the casino task, participants were asked to label stimuli as old (they had seen them in the casino task) or new.

Choices reflect reward history, novelty, and uncertainty

We first examined how reward history, uncertainty, and novelty influence choice. To investigate the influence of reward history, choice was modeled as a function of the difference in expected value between the left and right options ($EV_L - EV_R$), where the expected value was defined as the mean of a Beta distribution specified according to each option's reward history. As illustrated in Figure 2A, choice was robustly governed by reward history, and participants reliably sampled more rewarding options ($\beta_1 = 4.76, p = 3.1e-14$).

We probed the influence of novelty by focusing on the subset of trials in which a novel stimulus was offered. The proportion of trials in which the novel option was sampled as a function of the alternative option's expected value is illustrated in Figure 2B. Again, we observe a significantly positive slope, indicating that reward history exerts a strong force on choice ($\beta_1 = 4.9, p = 8.6e-14$), but we also observed a bias toward sampling the option with a novel stimulus on average ($\beta_0 = 0.36$, $p = 3.1e-6$).

We revisited this analysis to investigate the effect of uncertainty, which we quantified as the number of times a particular option has been sampled within a given block (see Figure 2C). An analysis of the proportion of trials in which the more uncertain option was sampled revealed a significantly positive slope, indicating that choice is strongly shaped by reward history ($\beta_1 = 4.1, p = 4.99e-10$), but we also observed an aversion to more uncertain options on average ($\beta_0 = -0.26, p = 4.18e-8$), indicating that in contrast to a novelty seeking bias, participants tend to shy away from more uncertain options.

The influence of reward history, novelty, and uncertainty adapted to task horizon

Our experimental design allowed us to probe how the influence of reward, novelty, and uncertainty adapted to the task horizon. We characterized these effects using a computationally agnostic logistic regression model to quantify the interaction between stimulus features (reward, novelty, and uncertainty) and trial number within each block (see Equation 2 in STAR Methods for details).

As illustrated in Figure 2D, reward history is a strong predictor of choice; participants preferring the option with a higher expected value ($\beta_{EV} = 1.43, p = 9.90e-14$). However, the influence of expected value diminished as participants progressed through a block of trials ($\beta_{EV:t} = -0.62, p = 1.19e-05$), suggesting that participants deviated from optimal outcome integration or, potentially, adopted a less exploitative strategy toward the end of each learning context (see Figure S1). Uncertainty also had a significant effect on choice, although participants expressed differing strategies at the start of a learning block; there was a roughly even split between uncertainty-seeking and -aversive

Please cite this article in press as: Cockburn et al., Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.05.025

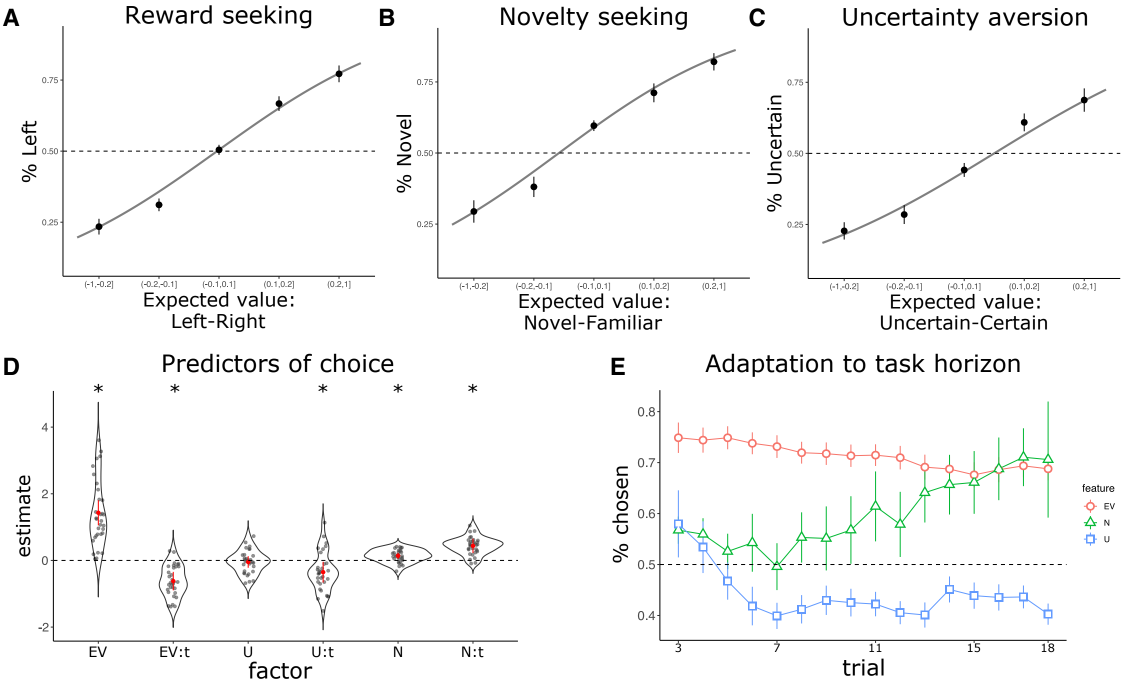

Figure 2. Behavioral evidence for the influence of reward history, novelty, and uncertainty on choice

(A) The proportion of trials in which the left option was chosen as a function of the difference in expected value between the left and right options across all trials.

(B) The proportion of trials in which the novel option was chosen as a function of the difference in expected value between the novel and familiar option across trials where a novel stimulus (offered < 3 times) was paired with a familiar stimulus.

(C) The proportion of trials in which the uncertain option was chosen as a function of the difference in expected value between familiar stimuli with differing uncertainty. Bin boundaries were defined to allow for an approximately equal number of trials per bin. Gray curves depict the choice proportions predicted by the statistical models fit to the data across an interval of values spaced to mirror distances between the plotted bins.

(D) Estimates from a logistic regression showing the influence of stimulus features on choice and their trajectory across a block of trials. Stars denote significant variable effects, gray points depict random effects, and red points and error bars depict fixed effects and 95% confidence intervals, respectively.

(E) The proportion of trials in which the option with a higher expected value was selected (red), the option that was most novel was selected (green), and the option that was most uncertain was selected (blue) across the task horizon within a sliding window of three trials. Points in (A), (B), (C), and (E) represent mean proportions across participants, and error bars depict the standard error of the participant mean scores.

individuals, resulting in a group average that did not differ from zero ($\beta_U = -0.04, p = 0.6$). However, a model that included uncertainty offered a significant improvement over a model that did not $\left(\chi^2(23, 32) = 210, p = 6.362\,e - 05\right)$, showing that uncertainty played a significant but varied role across individuals early in learning. In line with optimal theories of exploration (Gittins, 1979), participants grew increasingly reluctant to sample more uncertain options as task termination approached ($\beta_{U:t} = -0.37, p = 0.02$). Lastly, participants expressed a robust novelty-seeking bias at the start of the learning context ($\beta_N = 0.14, p = 0.003$). In contrast to the growing uncertainty aversion, novel options grew increasingly attractive as the block of trials unfolded ($\beta_{N:t} = 0.44, p = 7.42\,e - 05$).

To better illustrate these feature-specific trajectories, we extracted three subsets of trials: trials in which the two options being offered were both familiar but had different expected values (Figure 2E, red); trials in which both options were familiar but had different levels of uncertainty (blue); and trials in which one of the options was novel (green). Using a sliding window of three trials, Figure 2E illustrates the mean proportion of trials in which the option with higher expected value (red), higher uncertainty (blue), or higher novelty (green) was chosen as participants progressed through the learning contexts. Reflecting regression results, participants grew less likely to probe the more uncertain stimulus offered but also expressed an increasing preference for novel options despite their inherent uncertainty. We note that growing preference for novel options does not appear to be an effect of boredom, as demonstrated by a worse fit from a model that includes block number as an interaction term (see STAR Methods'"analysis of nuisance task variables").

Replication of novelty seeking and uncertainty aversion

We sought to establish the reliability of the key behavioral signatures highlighted in our fMRI study. Behavioral data were collected from n = 79 participants at the University of Toronto, where participants were exposed to a variant of the experimental protocol in which the first 6 blocks of the experiment replicate the task design described here (see STAR Methods for details).

Replicating our original findings, participants in this second study were influenced by reward history, stimulus novelty, and estimation uncertainty. Mirroring the analyses reported in Figures 2A–2C, participants exhibited a reliable preference for the option with a higher expected value ($\beta_1 = 3.77, p = 2\,e - 16$). Participants also elicited a robust novelty-seeking bias

$(\beta_1 = 0.23, \mathrm{p} = 1.26\,e-4)$ and uncertainty aversion on average $(\beta_0 = -0.16, \mathrm{p} = 5.56\,e-4)$. Using the same logistic-regression analysis illustrated in Figure 2D, we observed similar results to those of the original fMRI study. Participants were strongly influenced by reward history, and there was a waning effect of optimal reward integration across trials $(\beta_{EV} = 0.70, \mathrm{p} < 2\,e - 16;\ \beta_{EV:t} = -0.33, \mathrm{p} = 0.006)$. Novelty was not found to significantly bias choice early in a learning block but grew significantly more enticing across the block of trials $(\beta_N = 0.03, \mathrm{p} = 0.18;\ \beta_{N:t} = 0.14, \mathrm{p} < 0.01)$. We note that this same pattern is observed in our original fMRI study when this analyses was limited to only the first 6 blocks of learning $(\beta_N = 0.007, \mathrm{p} = 0.9;\ \beta_{N:t} = 0.68, \mathrm{p} < 0.0005)$. Finally, uncertainty did not consistently bias choice early in a learning block, but participants developed a growing aversion to more uncertain options as the task horizon approached $(\beta_U = -0.05, p = 0.16, \beta_{U:t} = -0.11,\ p = 0.04)$, demonstrating that participants in the replication study exhibited the same growing tension between novelty and uncertainty.

In summary, using a computationally agnostic regression model of choice, we found evidence of both exploration and exploitation during our task, and this evidence was replicated in a larger behavioral study. Of note, we identified clear conflict between features associated with exploration and their trajectories across the task at hand; this conflict is expressed as a growing distaste for uncertainty concomitant with a growing appetite for novel alternatives despite their inherent uncertainty. Using these results to benchmark and constrain our consideration of the mechanisms driving choice, we turned to computational models of learning and decision making to elucidate how these patterns emerge. We focused not only on how the trade-off between exploration and exploitation is regulated but also on the puzzling tension between novelty and uncertainty guiding exploration.

Computational model of choice

Regression analyses identified reward history, stimulus novelty, estimation uncertainty, and task horizon as features influencing choice. Here, we apply computational models of learning and decision making to behavior collected during the fMRI study to better understand how the explore/exploit trade-off is regulated in the human brain.

The task was modeled from the perspective of a forgetful Bayesian learner. In brief, each option is represented as a Beta distribution, which was defined according to a recency-weighted integration of previously observed outcomes, where novelty was accommodated by way of optimistic value initialization. From this representation we employ the distribution's mean and variance as the option's expected value and uncertainty, respectively. In line with theoretical and empirical results, uncertainty was incorporated into the decision-making process as a policy-biasing term (see STAR Methods for details).

Uncertainty and novelty were both shown to influence choice, but each followed separable trajectories across the task's horizon (see Figures 2D and 2E). We accommodated this adaptive influence into the model by including feature weights that consider progress through the learning context. Using free parameters that define the initial $(U_I\ , N_I)$ and terminal $(U_T\ , N_T)$ feature weights, the model can flexibly adjust how both uncertainty and novelty factor into the subjective utility of a given stimulus according to a linear trajectory across the task horizon.

As noted previously, participants expressed a growing distaste for uncertain options and an increasing preference for novel options despite their inherent uncertainty, presenting a tension that demands further scrutiny. This pattern of response could emerge from a system that considers both novelty and uncertainty as pertinent stimulus features but increases the drive to seek novelty at a rate that outpaces uncertainty aversion. However, we know of neither empirical evidence nor normative theory suggesting that novelty should be increasingly valued as a task approaches termination. Other explanations, such as boredom, do not offer a coherent explanation of the phenomena either; there was no effect of block number $(\beta_b = -0.08, p = 0.44)$, nor was there a significant interaction between block number and novelty $(\beta_{N:b} = -0.05, \mathrm{p} = 0.6)$, demonstrating that novelty-seeking strategies are not accounted for by variation in experiment duration (as opposed to block level) (see STAR Methods).

Alternatively, we reasoned that this pattern of behavior could emerge by way of an interaction between novelty and uncertainty processing in the brain; specifically, a system in which stimulus novelty interferes with the potency of uncertainty. In this framework, familiar options derive their subjective utility according to both reward history and an uncertainty bias. Conversely, visually novel options are valued according to their optimistically initialized expected value alone, absent consideration of the uncertainty bias. Under this scheme, as depicted in Figure 3A, an increasing reluctance to sample familiar uncertain options will push favor toward novel alternatives for which uncertainty is ignored, resulting in a growing propensity to sample novel options as the aversion to uncertainty gains strength.

To formalize this hypothesis as a computational model, we began with a baseline forgetful Bayesian reinforcement learning model that included two free parameters: a forgetting rate (η) that determined the rate of behavioral adaptation in response to observed outcomes; and a Softmax choice stochasticity parameter (β), which determined the degree to which choice relied on value. We augmented this baseline model to embody our hypothesized mechanisms of human exploration in what we label the familiarity-modulated upper-confidence-bound model (fmUCB). This model includes free parameters U_I and U_T to incorporate an adaptive uncertainty bias, and $N_I = N_T$ to facilitate a consistent value initialization bias throughout the task. Importantly, we address the tension between novelty and uncertainty by incorporating stimulus novelty as a modulatory mechanism governing the uncertainty bias, effectively blocking the influence of uncertainty when stimulus novelty is high.

Model characteristics and comparisons on behavioral data

We first examined the degree to which the fmUCB model faithfully reproduced the response patterns observed in participants' behavior. Having fit the model's free parameters to the behavioral data, in Figure 3C we illustrate the estimated effects identified in our participant group (in red along the x-axis) alongside

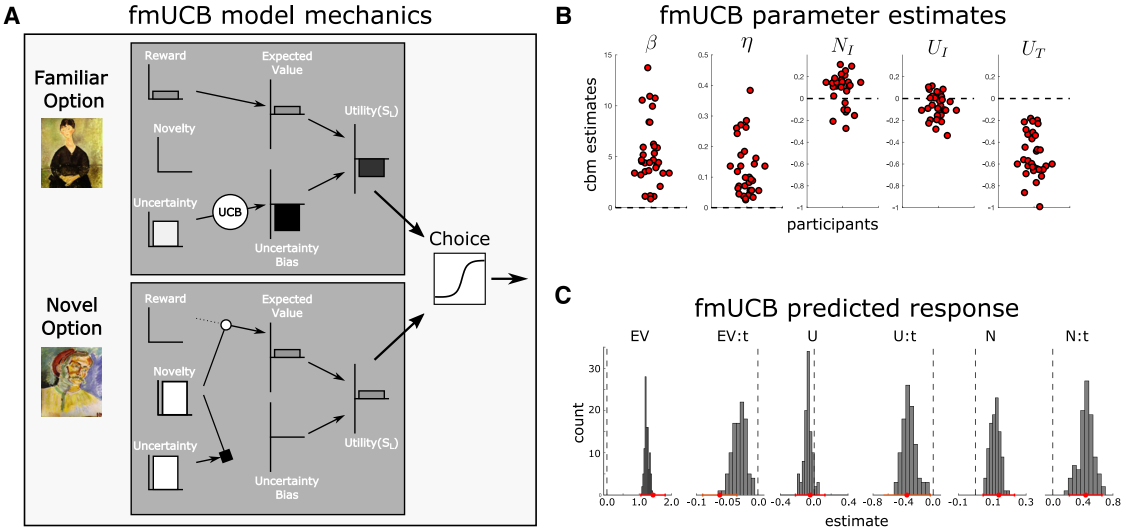

Figure 3. Computational modeling

(A) An illustration of the mechanisms driving choice in the fmUCB model when novel and familiar options were offered near the end of a learning context. The familiar option's utility combines the option's reward history and uncertainty bias, resulting in a negative net utility as a result of the overwhelming uncertainty aversion induced by the agent's proximity to task termination. Despite not having been sampled, the novel option is endowed with a positive expected value, and uncertainty is prevented from being integrated into the net utility (a black square depicts novelty's inhibition of the uncertainty bias).

(B) Parameter estimates for the fmUCB model (Softmax β governing choice determinacy, forgetting rate η, constant novelty shaping bias $N_I = N_T$, uncertainty bias intercept and terminal values U_I, U_T).

(C) Posterior predictive check depicting the degree to which the fmUCB model captures the response patterns highlighted by the computationally agnostic regression analysis. Histogram bars denote regression coefficient counts. Behavioral regression coefficient and 95% confidence intervals are depicted as points and bars in red along the x-axis.

the distribution of effects generated by the model. The fmUCB model recapitulates the effects observed in the behavioral sample well, and effects of novelty and uncertainty are faithfully reproduced, indicating that the fmUCB model does indeed capture critical patterns of interest in the behavioral data.

To test whether the fmUCB model offered a parsimonious explanation of participants' behavioral data, we performed a model comparison in which it was contrasted with alternative models. This alternative set of models included the above baseline model (including β and η as free parameters), as well as a family of models containing a full permutation of exploration-related variables without the additional constraints of the familiarity gating mechanism, defined by a parameter set that includes initial and terminal novelty initialization bias (N_I and N_T) as well as initial and terminal uncertainty bias weights (U_I, and U_T). We performed model comparison on the behavioral data from the fMRI dataset because it had sufficient numbers of trials per participant to enable model fitting and comparison. Using the CBM toolkit (Piray et al., 2019) to perform this comparison,we found that, with 96% exceedance probability, the fmUCB model offered the best explanation of the behavioral data given the set of models being considered.

As depicted in Figure 3B, the fmUCB model's optimized parameter estimates exhibit a positive estimate of N_I, indicating that expected values are indeed optimistically initialized (t(31) = 3.56, CI = [0.04, 0.14], p = 0.001). Furthermore, participants also manifested a decreasing uncertainty bias, consistent with the hypothesis that the uncertainty bias encapsulates the prospective benefit of uncertainty reduction ($U_I - U_T$: t(31) = 12.38, p = $1.57e-13$, CI: [0.27, 0.38]). Given that model comparison identified the fmUCB model as the best explanation of the behavioral data and that this model includes N_I, U_I, and U_T as free parameters, these results demonstrate that response patterns observed in the participant behavior are not a by-product of sampling biases (e.g., sampling more richly rewarded options at the expense of more uncertain and lower-valued options).

We next considered critical details about how the fmUCB model could be implemented. First, a reasonable alternative to optimistic initialization could incorporate novelty during the decision-making process as a separable bonus term (as is operationalized through the uncertainty bias term). Thus, we sought to probe for evidence supportive of either an optimistic initialization or a policy-biasing mechanism driving novelty seeking behavior. Second, the fmUCB model does not specify the mechanism by which novelty dampens the uncertainty bias. This pattern of response could emerge incidentally, where novel stimuli promote a decision before the bonus can be fully integrated into the subjective utility-guiding choice, or alternatively, the neural processes engaged by novel stimuli could directly interfere with the processes required to compute the uncertainty bias itself. Although these mechanisms are behaviorally indistinguishable, the profile of the underlying computational variables as they evolve over time differs between the implementations. Given that these variables should be encoded in the brain according to the model variant that is actually being utilized at the neural level, we aimed to determine whether we could distinguish between these implementational forms on the basis of the fMRI data.

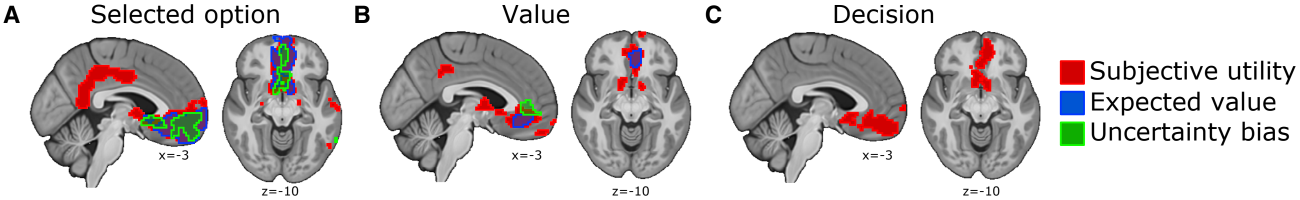

Figure 4. Neural correlates of computational variables guiding choice

(A) Neural correlates associated with the chosen option. Clusters including posterior cingulate gyrus and a large cluster extending from vmPFC to the ventral striatum were positively correlated with the subjective utility of the chosen stimulus (red). A cluster including vmPFC and the ventral striatum was positively correlated with the expected value (blue), and a cluster incorporating vmPFC was positively correlated with the uncertainty bias (green).

(B) Value contrasts (selected + rejected option value) exposed a cluster extending from vmPFC to the ventral striatum as positively correlated with the mean subjective utility of both options (red). A cluster in vmPFC was positively correlated with the expected value (blue), and a cluster in mPFC was correlated with the uncertainty bias (green).

(C) Decision contrasts (selected − rejected) identified a significant cluster associated with subjective utility extending from vmPFC to ventral striatum. Whole-brain maps for the signal of interest were tested with a cluster-forming threshold of $p < 0.001$ uncorrected, followed by cluster-level FWE correction at $p < 0.05$.

Neural correlates of subjective utility and preference

Before investigating the neural implementation of the fmUCB model's mechanisms, we sought to identify brain regions correlating with the model's key variables and focused first on the subjective utility of the stimuli being offered. To do so, we defined a GLM that included the stimulus utility, which is defined as the summed expected value and uncertainty bias as used by the fmUCB model to guide choice, for both the selected and rejected options as stimulus-locked parametric modulators. We probed for the neural correlates of the chosen option's subjective utility (see Figure 4A, red), which identified a positively correlated cluster composed of vmPFC (peak voxel: x = −10, y = 41, z = −10; t = 6.4) and ventral striatum (peak voxel: x = 5, y = 16, z = −4; t = 7.86), as well as a cluster spanning posterior cingulate cortex (peak voxel: x = −5, y = −51, z = 10; t = 5.44). We sought to better characterize these correlates by expanding this analysis to probe for signals contributing to valuation and/or the decision-making process itself. Noting that the value of both options is correlated with their sum (selected + rejected), whereas a comparative decision-making process is correlated with their difference (selected - rejected) (Elber-Dorozko and Loewenstein, 2018), we constructed and analyzed the corresponding contrasts. The summed subjective utility contrast identified a cluster extending from vmPFC (peak voxel: x = −8, y = 36, z = −14; t = 5.3) to ventral striatum (peak voxel: x = 12, y = 6, z = −7; t = 5.39), and a small cluster in posterior cingulate cortex (peak voxel: x = −5, y = −41, z = 40; t = 4.36) (Figure 4B red). Focusing next on the decision-making process, we identified a vmPFC cluster (peak voxel: x = −10, y = 46, z = −12; t = 4.91) that was positively correlated with the relative difference in subjective utility (selected − rejected; Figure 4C red). Consistent with numerous previous studies (Bartra et al., 2013; Clithero and Rangel, 2014), these results highlight the vmPFC as playing a central role in the valuation and decision-making process.

The fmUCB model derives the subjective utility for each option as the sum of an optimistically initialized expected value and a familiarity-modulated uncertainty bias. We sought to determine whether both of these signals were represented in the brain and, if so, whether they are they co-located and how they came to influence the decision-making process. In pursuit of this, we specified a second GLM in which the subjective utility was decomposed into its constituent expected-value and uncertainty-bias terms for both the selected and rejected options. An analysis of BOLD signal change positively correlated with the selected option stressed a prominent role for vmPFC in tracking stimulus features pertinent to value and decision making (see Figure 4A blue and green), with clusters positively correlated with novelty-biased expected value (peak voxel: x = −8, y = 36, z = −10; t = 6.62) as well as a largely overlapping cluster positively correlated with the uncertainty bias (peak voxel: x = −2, y = 51, z = −4; t = 5.1). We then repeated the valuation (selected + rejected) and decision (selected − rejected) contrast analyses for both expected value and the uncertainty bias. The valuation contrast exposed a vmPFC cluster that positively correlated with optimistically initialized expected values (peak voxel: x = −8, y = 36, z = −14; t = 4.89, Figure 4B blue), and a slightly more dorsal cluster in medial PFC was positively correlated with the uncertainty bias (peak voxel: x = 8, y = 42, z = 8; t = 4.53, Figure 4B green). Given that each option's raw estimation uncertainty was also included in the GLM, correlates of the uncertainty bias can be argued to be tracking the potential utility ascribed to the option's uncertainty independent of uncertainty itself. No regions were found to correlate with the relative difference between the option's expected values or between their uncertainty-bias terms. These results suggest that stimulus features themselves are not directly compared; rather, an integrated subjective utility is formed and used to guide choice.

Neural evidence of optimistic initialization

As noted previously, visual novelty could be incorporated to guide choice through two candidate mechanisms: through optimistic initialization of an option's expected value as embodied by the fmUCB model or through a separable novelty-bias term integrated into the policy at the time of decision. Simulations showed that these two candidate mechanisms are not identifiable from behavior alone, from which the correct generative mechanism can only be identified with 54% accuracy (see STAR Methods). Crucially, although our task design does not support behavioral separability, these two mechanisms differ with respect to the expected value ascribed to novel options and their subsequent RPEs when chosen, and as such, we leverage fMRI activity to differentiate between these two implementations.

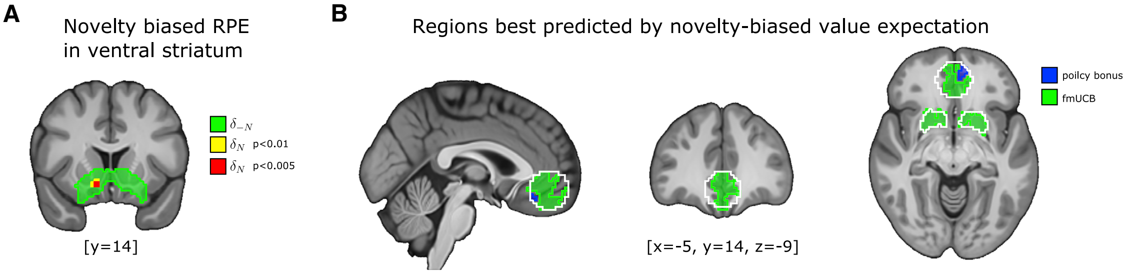

Figure 5. Representation of optimistic initialization in vmPFC and the ventral striatum

(A) Standard reward prediction errors are associated with robust activity in the ventral striatum (green). The additional component of the RPE attributed to optimistic initialization was also identified in the ventral striatum.

(B) Signal change in vmPFC and the ventral striatum is better explained by optimistic initialization than a linear bonus mechanism. Voxel-wise model comparison plots report voxel exceedance probability > 0.9, favoring the fmUCB model (green) or the linear bonus model (blue). The spatial mask encompassing vmPFC and the ventral striatum is bordered in white.

Optimistic initialization distorts value expectations and, transitively, the RPEs associated with novel options relative to familiar counterparts. This phenomenon has been exploited to identify biased signal change in ventral striatum unique to optimistic initialization (Wittmann et al., 2008). We replicated this analysis by extracting two RPE signals from the fmUCB model's timecourse; the first RPE (δ) represents the signal generated by the fmUCB model that includes the optimistic initialization component, and the second is the RPE as would be computed in the absence of any effect of novelty on value expectations $(\cdot\delta_{-N})$. From this we defined a regressor representing the novelty component of the RPE as $\delta_N = \delta - \delta_{-N}$. Consistent with previous findings (Daw et al., 2006; O'Doherty et al., 2003), we found that a GLM that included both δ_{-N} and δ_N as feedback-locked parametric regressors identified a strong correlation between the standard RPE (δ_{-N}) and activation in ventral striatum (see Figure 5A). Importantly, voxels in the right ventral striatum also correlated with the novelty-biased RPE component encoded by δ_N above and beyond the correlation found with the basic RPE signal (peak voxel: $x = 10$, $y = 16$, $z = -9$; $t = 3.02$, $p < 0.005$ height threshold; $p = 0.05$ SVC), supporting the presence of a novelty-biased RPE signal and consistent with (and overlapping spatially with) previous findings reported by Wittmann et al. (2008).

Pursuing this line of investigation further, we compared the variance explained by the optimistic initialization and policy-bias mechanisms by using Bayesian model comparison in target ROIs. To do so, we fit both models to the behavioral data (see Equation 16 for the policy-bias model) and extracted the model-predicted timecourse for the chosen option's expected value and the corresponding RPE. We estimated first-level GLMs for each participant by using both computational models and compared maps by using SPM's Bayesian model selection. We included the vmPFC as a region associated with valuation and choice (Clithero and Rangel, 2014; Bartra et al., 2013) and defined by a 15 mm radius sphere centered on peak voxel coordinates identified by (Clithero and Rangel, 2014), and we included ventral striatum as a second ROI given its association with reward prediction errors as found above (Wittmann et al., 2008; O'Doherty et al., 2003). As depicted in Figures 5B and 5A, comparison of the variance explained by either model shows that 421 voxels in vmPFC (56% of 746 voxels) and 357 voxels in ventral striatum (95% of 376 voxels) were best explained by the optimistic initialization mechanism (exceedance threshold $\geq$ 0.9, cluster size $\geq$ 10). Conversely, only 94 voxels were best explained by the policy-bias mechanism across both regions, suggesting that optimistic value initiation provides a better overall account for expected value and RPE signals in both vmPFC and ventral striatum. We note that value and RPE signals derived from both models yield significant effects in overlapping brain areas, and these regions are encompassed by our ROIs. Thus, our conclusions about model comparison are not merely an artifact of the ROIs selected because these ROIs capture the key regions in which statistically significant fMRI correlates of these variables are present according to either implementation (see Figure S5).

Novelty disrupts the computation of the uncertainty bias

The fmUCB model adapts the uncertainty bias according to the task's horizon, growing increasingly unwilling to probe uncertain familiar options toward the end of a learning context (see Figures 2D and 2E). The model also expresses a growing preference for novel options despite their inherent uncertainty, which is accommodated through a familiarity modulated uncertainty bias. Here, we query the fMRI data for patterns consistent with two plausible candidate implementations of the model's uncertainty bias.

As previously outlined, this pattern of response could emerge incidentally if novelty promoted choice before the uncertainty bias could be fully integrated into the subjective utility. Alternatively, the neural processes required for computing the uncertainty bias might be directly obstructed by processes otherwise engaged by novel stimuli. These two competing implementations cannot be distinguished on the basis of the behavioral data alone because they yield equivalent choice behavior. Therefore, we sought to utilize the constraints imposed by their neural implementation to differentiate between them. Specifically, direct interference predicts the absence of uncertainty-bias signals associated with novel stimuli, whereas incidental choice induction implies that the cognitive processes responsible for computing the uncertainty bias ought to be present despite their inability to guide behavior.

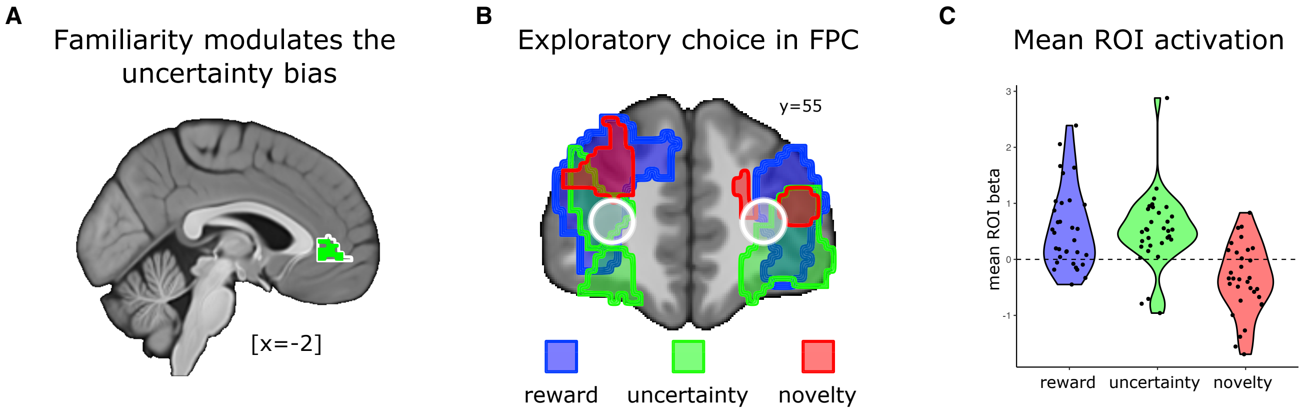

Figure 6. Effects of novelty on uncertainty bias and choice

(A) Model comparison in mPFC supports a novelty-induced interference implementation of the familiarity-modulated uncertainty bias. Voxel-wise model comparison shows that the exceedance probability favors the explicit interference implementation of the fmUCB model. Masked brain plots show voxels favoring the explicit fmUCB model with an exceedance probability > 0.9 in green (mask encompassing mPFC bordered in white).

(B) Significant clusters in bi-lateral FPC are positively correlated with choosing the lower valued option (blue) and the more uncertain option (green) and negatively correlated with choosing the more novel option (red). Brain maps reported exceed a cluster-forming threshold of $p < 0.001$ uncorrected, followed by cluster-level FWE correction at $p < 0.05$.

(C) ROI analysis of mean beta estimates in ROIs situated at bilateral FPC (white bordered circles); analysis is taken from Daw et al., 2006 according to choice.

We used a model-comparison approach to determine which of these implementations better describe the pattern of activity in the brain as a post-hoc analysis. Our previous analyses show that the uncertainty bias is reflected in mPFC (see Figure 4B), highlighting this as a target ROI within which to evaluate the predictions of the two mechanisms. We compared a GLM specified according to the timecourse of variables extracted from the fmUCB model to a GLM specified according to variables from a model in which the uncertainty bias was not modulated by stimulus novelty. As illustrated in Figure 6A, a cluster of 79 voxels was found to be better represented by the familiarity-modulated bias term, whereas no cluster favored the uncertainty bias in the absence of familiarity modulation (exceedance probability > 0.9, cluster threshold = 10). We note that this analysis is not independent of the model-selection criteria. Importantly, although no significant clusters were found to correlate with the unmodulated uncertainty bias, peak voxels were identified in this same ROI at a more liberal threshold, suggesting that this ROI offers the best opportunity to detect signals associated with the alternative model should they be present (see Figure S6). Further demonstrating the suitability of the ROI for comparison, model comparison targeting critical trials of differentiation between the two mechanisms across a broad sphere in mPFC favor the familiarity-gated mechanism more strongly (see supplemental information). This suggests that the brain's uncertainty bias is diminished when processing novel stimuli and is consistent with a mechanism in which novelty actively inhibits the processes through which uncertainty-directed decision making is guided.

Ultimately, the computational variables and cognitive processes under investigation produce a behavioral choice. Previous findings have shown greater FPC engagement when indivudals forgo the most rewarding option in order to explore a lower-value alternative (Daw et al., 2006; Boorman et al., 2009) and when choice targets uncertainty reduction (Badre et al., 2012; Zajkowski et al., 2017). We leverage the fact that our experimental design offered options that varied in terms of expected value, uncertainty, and novelty to probe the activation patterns across these dimensions according to the choice that was made.

We constructed a GLM that included three response-locked parametric regressors indicating whether the option with lower expected value was selected (random exploration), whether the option with higher estimation uncertainty was sampled (uncertainty-directed exploration), and whether a novel option was chosen (novelty-driven exploration). As illustrated in Figure 6B, this analysis revealed differential FPC engagement across choices. We identified bilateral FPC clusters that showed elevated activation when either the lower-valued option was chosen (rFPC: x = 32, y = 51, z = 20, t = 5.94; lFPC: x = −27, y = 51, z = 15, t = 5.4) or the option with higher uncertainty was sampled (rFPC: x = 25, y = 56, z = −12, t = 5.96; lFPC: x = −25, y = 64, z = 0, t = 4.49). In contrast, we observed a significant bilateral reduction in FPC activity when novel stimuli were chosen (rFPC: x = 22 ,y = 51, z = 23, t = 4.91; lFPC: x = −20, y = 46.5, z = 13, t = 5.02). These effects were corroborated by an ROI analysis centered on FPC coordinates previously reported to be associated with exploration (Daw et al., 2006) (Figures 6B and 6C); both value- (t(31) = 3.27, p = 0.003) and uncertainty- (t(31) = 4.3, p = 1.45 *e* − 04) driven exploration associated with greater FPC activation. However, sampling novel options elicited a significantly reduced activation in those same regions (t(31) = −3.36, p = 0.002).

DISCUSSION

In this study we investigate how the human brain balances a fundamental tension between exploration and exploitation.

Previous studies have characterized a range of normative and heuristic solutions to the problem (Daw et al., 2006; Wilson et al., 2014; Badre et al., 2012; Frank et al., 2009; Blanchard and Gershman, 2018; Wittmann et al., 2008; Gershman, 2018; Payzan-LeNestour et al., 2013; Domenech et al., 2020; Costa et al., 2019). However, novelty and uncertainty, two key variables known to influence choice, had yet to be simultaneously probed. Here, we show that humans adaptively evaluate the benefit of reducing uncertainty and that they grow less likely to target uncertain options as the prospective advantage of additional information declines. In contrast, despite the inherent uncertainty of novel options, they were pursued irrespective of the task horizon. Notably, these findings were also replicated in a larger behavioral sample, demonstrating generalizability across populations.

Using a computational model of choice constrained by neural data, we demonstrate how this apparent tension between uncertainty-guided choice and novelty seeking might be resolved in the human brain. We found that choice reflected an adaptive uncertainty bias tracking the prospective value of information, which was reflected by BOLD signal change in mPFC. Novelty was found to bias choice by corrupting reward expectations, as reflected by BOLD signal change in ventral striatum and vmPFC, resulting in persistent novelty-seeking behavior. In addition to inflating an option's expected value, visual novelty also diminished the influence of uncertainty on choice, where a signal change in mPFC reflected a familiarity-modulated uncertainty bias. This, we argue, is consistent with an antagonistic interaction between novelty and uncertainty processing. Thus, we propose that the increasing appetite for visually novel options reflects two separable processes that act to promote exploration: firstly, inflated reward expectations guide the brain's exploitative circuitry toward sampling novel options, and second, inhibited uncertainty processing diminishes the otherwise aversive nature of the unknown when new information has low prospective utility.

Our results also build on previous work exposing the role of the vmPFC as tracking both the probability of choice (Daw et al., 2006) and the relative difference between the selected and rejected values (Boorman et al., 2009), implicating vmPFC in both valuation and decision making (Clithero and Rangel, 2014). In line with these reports, we found that both the subjective utility of the options under consideration and their relative difference was represented by vmPFC (Figures 4B and 4C). Recent studies have also suggested that vmPFC contributes to the regulation of exploration and exploitation by monitoring outcomes, signalling the degree to which predictions are being met (Domenech et al., 2020), and others have reported vmPFC signal consistent with the value ascribed to information seeking (Trudel et al., 2020). Our results build on these findings and others (Suzuki et al., 2017) showing that vmPFC plays a role in integrating stimulus feature values, in this case, to regulate the trade-off between exploration and exploitation. Consistent with previous work arguing that vmPFC acts as a value integration hub where disparate stimulus features are weighted according to current goals (Hare et al., 2009; Suzuki et al., 2017; O'Doherty et al., 2017), we observed a robust vmPFC signal reflecting the relative difference in the subjective utility of options being considered. This suggests that the expected value of reward and the expected value of information are integrated to form a subjective stimulus value from which behavior may be directed.

Previous research examining the neural correlates of human exploration have implicated FPC; however, the nature of its role remains elusive. FPC has been shown to be engaged when participants opted to forgo the most rewarding option (Daw et al., 2006) and in tracking the relative uncertainty of the available options (Badre et al., 2012; Yoshida and Ishii, 2006), implicating this region with both random and directed exploration. Consistent with these findings, we found that FPC was preferentially engaged when participants decided to sample either the lower-valued or more uncertain option (Figures 6B and 6C). However, we also found that FPC was suppressed when novel options were chosen. This, we argue, offers further support for the hypothesis that the brain does not frame novelty seeking as exploration per se, but as value exploitation rooted in biased value expectations.

The present study also has methodological implications that go beyond the specific research question in that it offers a demonstration of how neural evidence can be used to constrain and arbitrate between different computational mechanisms. This demonstration is pertinent to a persistent debate in the psychological literature about the utility (or absence thereof) of neuroscience measures. Utilizing a model comparison approach on fMRI data, we were able to obtain evidence in support of one particular model structure, thereby validating the importance of using brain measures alongside behavioral ones to advance theoretical understanding of cognitive processes (Palmeri et al., 2017; Henson, 2005; Page, 2006; Collins and Cockburn, 2020).

Questions remain about the benefit of this particular configuration, in which stimulus novelty parasitizes the brain's reward circuitry to promote exploration while uncertainty-directed sampling relies on separable integrated feature values. We speculate that this offers a parsimonious decomposition of an otherwise intractable decision problem. Inferring the prospective benefit of strategically reducing uncertainty can offer significant benefit, particularly in volatile environments where reward contingencies and goals can change rapidly. However, the computational demands of this approach are high, and in unfamiliar circumstances, the computations are most likely nothing more than guess work. We suggest that by diverting computational resources away from ill-suited strategies in favor of more computationally efficient heuristics such as optimistic initialization, the brain can begin to bridge the gap from a computational intractable scenario toward a manageable representation of the landscape where more adaptive strategies such as uncertainty-driven exploration can be beneficially applied.

To conclude, in this study we offer further insight into how the human brain balances the explore/exploit trade-off. By systematically decoupling stimulus novelty and uncertainty, and by leveraging neural data to constrain models of human learning and decision making, we show that human exploration simultaneously targets different stimulus features by involving distinct strategies with potentially conflicting preferences. Future research investigating how novelty and uncertainty interact in different contexts will offer a more refined understanding of how the brain balances the explore/exploit trade-off, as will more sophisticated methods of characterizing how the brain maintains and leverages representations of uncertainty, how the brain computes the value of information gain, and the mechanisms through which stimulus novelty is transformed from the visual domain to that of value. How the brain resolves these tensions has a significant impact on behavior, highlighting potential avenues through which dysregulation of the balance between exploring new alternatives versus staying the course may be investigated.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCE TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Participants: fMRI study
  - Participants: Behavioral replication study
- METHOD DETAILS
  - Experimental design: fMRI study
  - Experimental design: Behavioral replication study
  - Software
  - fMRI data acquisition
  - fMRI data preprocessing and analysis
  - GLM design for fMRI analysis
  - Neural correlates of expected reward associated with the linear bonus model
  - Neural correlates of the uncertainty bias absent familiarity modulation
  - Targeted analyses reveal robust novelty disruption of uncertainty bias processes
  - Regions of interest and small volume correction
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Behavioral analysis
  - Analysis of task comprehension
  - Analysis of nuisance task variables
  - Computational modeling of behavior
  - Parameter estimation and model comparison
  - Behavioral confusability analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2022.05.025.

ACKNOWLEDGMENTS

This work was funded by grants from the US National Science Foundation (1207573) to J.O.D., the Natural Science and Engineering Research Council (RGPIN-2018-05946) to W.C., and the Social Sciences and Humanities Research Council (CGS-MSFSS) to V.M.

AUTHOR CONTRIBUTIONS

J.C. and J.P.O. conceived of the experiments and data analysis. J.C., J.P.O., V.M., and W.C. wrote the manuscript. J.C. carried out the fMRI experiment, and V.M. carried out the replication study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 8, 2021
Revised: February 18, 2022
Accepted: May 27, 2022
Published: July 8, 2022

REFERENCES

Agrawal, R. (1995). Sample mean based index policies by O(log n) regret for the multi-armed bandit problem. Adv. Appl. Probab. *27*, 1054–1078. https://doi.org/10.1017/s0001867800047790.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Mach. Learn. *47*, 235–256. https://doi.org/10.1023/a:1013689704352.

Avants, B.B., Tustison, N., and Song, G. (2009). Advanced normalization tools (ants). Insight j *2*, 1–35.

Badre, D., Doll, B.B., Long, N.M., and Frank, M.J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron *73*, 595–607. https://doi.org/10.1016/j.neuron.2011.12.025.

Bartra, O., McGuire, J.T., and Kable, J.W. (2013). The valuation system: a coordinate-based meta-analysis of bold fmri experiments examining neural correlates of subjective value. Neuroimage *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1, pp. 1–8.

Raja Beharelle, A., Polanía, R., Hare, T.A., and Ruff, C.C. (2015). Transcranial stimulation over frontopolar cortex elucidates the choice attributes and neural mechanisms used to resolve exploration–exploitation trade-offs. J. Neurosci. *35*, 14544–14556. https://doi.org/10.1523/jneurosci.2322-15.2015.

Blanchard, T.C., and Gershman, S.J. (2018). Pure correlates of exploration and exploitation in the human brain. Cognit. Affect Behav. Neurosci. *18*, 117–126. https://doi.org/10.3758/s13415-017-0556-2.

Boorman, E.D., Behrens, T.E., Woolrich, M.W., and Rushworth, M.F. (2009). How green is the grass on the other side? frontopolar cortex and the evidence in favor of alternative courses of action. Neuron *62*, 733–743. https://doi.org/10.1016/j.neuron.2009.05.014.

Brafman, R.I., and Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. J. Mach. Learn. Res. *3*, 213–231.

Brainard, D.H., and Vision, S. (1997). The psychophysics toolbox. Spatial Vis. *10*, 433–436. https://doi.org/10.1163/156856897x00357.

Bunge, S.A., and Wendelken, C. (2009). Comparing the bird in the hand with the ones in the bush. Neuron *62*, 609–611. https://doi.org/10.1016/j.neuron.2009.05.020.

Clithero, J.A., and Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. Soc. Cognit. Affect Neurosci. *9*, 1289–1302. https://doi.org/10.1093/scan/nst106.

Cohen, J.D., McClure, S.M., and Yu, A.J. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. Phil. Trans. Biol. Sci. *362*, 933–942. https://doi.org/10.1098/rstb.2007.2098.

Collins, A.G.E., and Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. Nat. Rev. Neurosci. *21*, 576–586. https://doi.org/10.1038/s41583-020-0355-6.

Costa, V.D., Mitz, A.R., and Averbeck, B.B. (2019). Subcortical substrates of explore-exploit decisions in primates. Neuron *103*, 533–545.e5. https://doi.org/10.1016/j.neuron.2019.05.017.

Daffner, K.R., Mesulam, M.M., Scinto, L.F., Cohen, L.G., Kennedy, B.P., West, W.C., and Holcomb, P.J. (1998). Regulation of attention to novel stimuli by frontal lobes: an event-related potential study. Neuroreport *9*, 787–791. https://doi.org/10.1097/00001756-199803300-00004.

Daw, N.D., O'doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879. https://doi.org/10.1038/nature04766.

Domenech, P., Rheims, S., and Koechlin, E. (2020). Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. Science *369*, eabb0184. https://doi.org/10.1126/science.abb0184.

Elber-Dorozko, L., and Loewenstein, Y. (2018). Striatal action-value neurons reconsidered. Elife *7*, e34248. https://doi.org/10.7554/elife.34248.

Ennaceur, A., and Delacour, J. (1988). A new one-trial test for neurobiological studies of memory in rats. 1: behavioral data. Behav. Brain Res. *31*, 47–59. https://doi.org/10.1016/0166-4328(88)90157-x.

Fantz, R.L. (1964). Visual experience in infants: decreased attention to familiar patterns relative to novel ones. Science *146*, 668–670. https://doi.org/10.1126/science.146.3644.668.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat. Neurosci. *12*, 1062–1068. https://doi.org/10.1038/nn.2342.

Gershman, S.J. (2018). Deconstructing the human algorithms for exploration. Cognition *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014.

Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. J. Roy. Stat. Soc. B *41*, 148–164. https://doi.org/10.1111/j.2517-6161.1979.tb01068.x.

Hare, T.A., Camerer, C.F., Rangel, A., and RAngel, A. (2009). Self-control in decision-making involves modulation of the vmpfc valuation system. Science *47*, S95. https://doi.org/10.1016/s1053-8119(09)70776-1.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? Q. J. Exp. Psychol. *58*, 193–233. https://doi.org/10.1080/02724980443000502.

Horvitz, J.C., Stewart, T., and Jacobs, B.L. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. Brain Res. *759*, 251–258. https://doi.org/10.1016/s0006-8993(97)00265-5.

Hughes, R.N. (2007). Neotic preferences in laboratory rodents: issues, assessment and substrates. Neurosci. Biobehav. Rev. *31*, 441–464. https://doi.org/10.1016/j.neubiorev.2006.11.004.

Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. Neural Network. *15*, 549–559. https://doi.org/10.1016/s0893-6080(02)00048-5.

Katehakis, M.N., and Robbins, H. (1995). Sequential choice from several populations. Proc. Natl. Acad. Sci. USA *92*, 8584–8585. https://doi.org/10.1073/pnas.92.19.8584.

Kidd, C., Piantadosi, S.T., and Aslin, R.N. (2012). The goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. PLoS One *7*, e36399. https://doi.org/10.1371/journal.pone.0036399.

Kidd, C., Piantadosi, S.T., and Aslin, R.N. (2014). The goldilocks effect in infant auditory attention. Child Dev. *85*, 1795–1804. https://doi.org/10.1111/cdev.12263.

Krebs, R.M., Schott, B.H., Schütze, H., and Düzel, E. (2009). The novelty exploration bonus and its attentional modulation. Neuropsychologia *47*, 2272–2281. https://doi.org/10.1016/j.neuropsychologia.2009.01.015.

Lau, B., and Glimcher, P.W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. J. Exp. Anal. Behav. *84*, 555–579. https://doi.org/10.1901/jeab.2005.110-04.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. J. Neurosci. *16*, 1936–1947. https://doi.org/10.1523/jneurosci.16-05-01936.1996.

Ng, A.Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: theory and application to reward shaping. ICML *99*, 278–287.

O'Doherty, J.P., Cockburn, J., and Pauli, W.M. (2017). Learning, reward, and decision making. Annu. Rev. Psychol. *68*, 73–100. https://doi.org/10.1146/annurev-psych-010416-044216.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *38*, 329–337. https://doi.org/10.1016/s0896-6273(03)00169-7.

Page, M.P. (2006). What can't functional neuroimaging tell the cognitive psychologist? Cortex *42*, 428–443. https://doi.org/10.1016/s0010-9452(08)70375-7.

Palmeri, T.J., Love, B.C., and Turner, B.M. (2017). Model-based cognitive neuroscience. J. Math. Psychol. *76*, 59–64. https://doi.org/10.1016/j.jmp.2016.10.010.

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., and O'Doherty, J. (2013). The neural representation of unexpected uncertainty during value-based decision making. Neuron *79*, 191–201. https://doi.org/10.1016/j.neuron.2013.04.037.

Pelli, D.G., and Vision, S. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. Spatial Vis. *10*, 437–442. https://doi.org/10.1163/156856897x00366.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., and Nichols, T.E. (2011). Statistical Parametric Mapping: The Analysis of Functional Brain Images (Elsevier).

Piray, P., Dezfouli, A., Heskes, T., Frank, M.J., and Daw, N.D. (2019). Hierarchical bayesian inference for concurrent model fitting and comparison for group studies. PLoS Comput. Biol. *15*, e1007043. https://doi.org/10.1371/journal.pcbi.1007043.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J. Neurophysiol. *80*, 1–27. https://doi.org/10.1152/jn.1998.80.1.1.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., et al. (2004). Advances in functional and structural mr image analysis and implementation as fsl. Neuroimage *23*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051.

Suzuki, S., Cross, L., and O'Doherty, J.P. (2017). Elucidating the underlying components of food valuation in the human orbitofrontal cortex. Nat. Neurosci. *20*, 1780–1786. https://doi.org/10.1038/s41593-017-0008-x.

Trudel, N., Scholl, J., Klein-Flügge, M.C., Fouragnan, E., Tankelevitch, L., Wittmann, M.K., and Rushworth, M.F.S. (2020). Polarity of uncertainty representation during exploration and exploitation in ventromedial prefrontal cortex. Nat. Human Behav. *5*, 83–98. https://doi.org/10.1038/s41562-020-0929-3.

Tyszka, J.M., and Pauli, W.M. (2016). In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. Hum. Brain Mapp. *37*, 3979–3998. https://doi.org/10.1002/hbm.23289.

Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A., and Cohen, J.D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. J. Exp. Psychol. Gen. *143*, 2074–2081. https://doi.org/10.1037/a0038199.

Wittmann, B.C., Daw, N.D., Seymour, B., and Dolan, R.J. (2008). Striatal activity underlies novelty-based choice in humans. Neuron *58*, 967–973. https://doi.org/10.1016/j.neuron.2008.04.027.

Yoshida, W., and Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. Neuron *50*, 781–789. https://doi.org/10.1016/j.neuron.2006.05.006.

Zajkowski, W.K., Kossut, M., and Wilson, R.C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. Elife *6*, e27430. https://doi.org/10.7554/elife.27430.

STAR★METHODS

KEY RESOURCE TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Behavioral data	This paper	https://osf.io/asqfd/
fMRI data	This paper	https://neurovault.org/collections/HUYXCWBV/
Software and algorithms		
MATLAB R2018a	MathWorks	https://www.mathworks.com/
FSL v.6.0	Smith et al. (2004)	https://fsl.fmrib.ox.ac.uk/ fsl/fslwiki
SPM 12	Penny et al. (2011)	https://www.fil.ion.ucl.ac.uk/spm/software/spm12/
ANTs	Avants et al. (2009)	http://stnava.github.io/ANTs/
Psychtoolbox 3.0	Brainard & Vision (1997)	http://psychtoolbox.org
R Studio (R 4.0.2)	RStudio Team	https://rstudio.com/
lme4	Bates et al. (2015)	https://cran.r-project.org/web/packages/lme4/index.html
Task and analysis code	This paper	https://zenodo.org/badge/latestdoi/487400076

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jeffrey Cockburn (jcockbur@caltech.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Group level t-maps have been deposited at Neurovault and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- Behavioral data have been deposited at OSF and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- Task and analysis code are publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants: fMRI study

We recruited 33 participants from the Pasadena community to take part in our study (age range [18–41], mean 27 years, 13 female). One participant was removed from the sample due to excess movement and poor performance (sleeping in the scanner), leaving a sample of 32 human participants. All participants were English speakers, had normal/corrected-to-normal vision, and had no history of neurological or psychiatric disease. Participants were paid a $40 base-rate plus a performance bonus ranging from $5-$15. The study was approved by the Caltech IRB and participants gave their informed consent to take part in the study.

Participants: Behavioral replication study

The replication study included 79 participants (age range [18–42], mean = 21 years, 48 female) from the Toronto community. All participants were required to be fluent English speakers and have normal or corrected-to-normal vision. 77 participants reported no prior history of psychiatric or neurological disease, but 2 additional participants reported a prior history of psychiatric illness. Those two participants are still included in the reported analyses because omitting them made no substantive difference to the results. Participants were paid a base rate of $10 per hour plus a performance bonus up to $20. All participants gave their informed consent to participate in the study in accordance with the Research Ethics Board at the University of Toronto.

METHOD DETAILS

Experimental design: fMRI study

Participants performed 20 blocks of a finite horizon multi-armed bandit task designed to expose the impact of reward, uncertainty and novelty in balancing the explore/exploit trade-off while undergoing four consecutive 15-min fMRI sessions (5 blocks per session). Both the task and the instructions were presented using Psychtoolbox-3 for Matlab. On each trial, participants were asked to choose between two slot machines. Having selected one of the machines, they were informed of the machines payout, either $1US (win) or $0US (loss). Participants were instructed that each block would consist 18 to 23 trials, and to encourage balanced attention and motivation throughout the task, they were informed that one block would be selected at random at the end of the experiment and they would be awarded the earning collected during that block as a cash bonus.

Each block was structured to include five visually unique and identifiable slot machines, three of which used familiar stimuli that had been seen during previous blocks, while the remaining two used novel stimuli that had not yet been shown (43 stimuli in total). The five slot machines that could be offered in a given block were each associated with a fixed probability winning, which was sampled from a linearly spaced range of [0.2–0.8]. Participants were informed that each slot machine had a fixed probability of winning within a block, but all machines were re-set at the start of each block, and as such, anything learned in previous blocks would not apply in blocks to come. Using sampling randomization, the two slot machines offered on each trial varied in terms of the number of previous exposures (novelty manipulation), as well as the number of times they'd been sampled in the current block (uncertainty manipulation), allowing us to to systematically examine the influence of novelty and uncertainty across the horizon of trials within a given block.

Following the slot machine task, participants performed a recognition test designed to probe their recall of which machines they had observed. They were asked to label 86 stimuli as 'old' or 'new', half of which had been used during the multi-armed bandit task, and half of which were not. Participants exhibited exceptional performance on the memory probe task, with mean accuracy of 89% (min 77%, max 97%), indicating the efficacy of the novelty manipulation.

We explicitly designed the task to discourage choice autocorrelations, and to be sensitive to response patterns in which choices were indifferent to stimulus reward history. We adopted a 5 choose 2 option sampling strategy to determine which stimuli were offered on each trial. This design implies that the stimulus sampled on trial t might not be available on trial $t+1$, and that both stimuli offered could be of high value or both could be of low value. This structure makes value-indifferent strategies such as sticky choice or win-stay/lose-shift difficult to employ consistently by forcing a more evenly distributed sampling than would occur if all options were simultaneously available. Additionally, by revaluing stimuli explicitly at the start of each block, and by randomizing the left/right presentation location across trials we could be sensitive to outcome insensitive action and stimulus choice autocorrelations (see STAR Methods: Computational modeling of behavior).

Experimental design: Behavioral replication study

The replication study consisted of an adapted version of the task described in the preceding section, with modifications described here. Prior to the start of the task, participants were instructed that they would be playing for points that would then be converted to a real monetary bonus up to a maximum of $20cnd. Because this version of the task included situations of monetary loss in affective conditions instantiated after an initial baseline condition, participants were initially endowed with a starting sum of 1,200 pts, with machines paying out 50 pts for a win and 0 pts otherwise. The baseline condition mirrors the design of the fMRI study; the affective manipulations only start after the baseline condition is complete and will be reported elsewhere.

Participants completed 6 blocks of the baseline condition consisting of 23 trials each. On each trial, participants were asked to chose between two slot machine from a set of six, each of which was associated with a fixed probability of winning, either sampled form a linearly spaced range of [0.2, 0.8] or from the set comprising [0.2, 0.44, 0.48, 0.52, 0.56, 0.6]. The assigned set of fixed win probabilities for a given block was chosen randomly, and participants were similarly instructed that the machines re-set at the start of each block. The structure of the novelty and uncertainty manipulations followed that reported in the fMRI study, though with two novel and four familiar stimuli. Two familiar stimuli were presented for the first 2 trials, with the first novel and third familiar stimuli introduced between trials 3–5. The remainder of the set (second novel and fourth familiar stimuli) were introduced between trial 8–19 in a pseudo-randomised manner. Relevant for the affective manipulation but independent of the primary multi-armed bandit task reported here, participants were probed with a subjective mood rating scale in 2 of the 6 blocks; these ratings are not analyzed further here.

As with the fMRI study, participants performed a recognition memory task following the multi-armed bandit task. In the replication sample participants similarly exhibited good recognition memory accuracy (mean = 82%, s.d. = 12%)

Software

Experiments were coded using Matlab, and presented using the Psychophysics Toolbox extensions (Brainard and Vision, 1997; Pelli and Vision, 1997). Behavioral analyses were conducted using the lme4 package in the R programming language for mixed-effect

Please cite this article in press as: Cockburn et al., Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.05.025

modeling Bates et al. (2015), and computational models were fit using the cmb toolkit (Piray et al., 2019). MRI data was analyzed using FSL, ANTs and SPM12.

fMRI data acquisition

Imaging data was collected at the Caltech Brain Imaging Center (Pasadena, CA) using a 3T Siemens Magneto TrioTim scanner using a 32-channel radio frequency coil. Functional scans were acquired using multiband acceleration of 4, 56 slices, voxel size = 2.5 mm isotropic, TR = 1,000 ms, TE = 30 ms, FA = 60 °, FOV = 200 mm x 200 mm. T1 and T2 weighted anatomical high-resolution scans were collected with 0.9 mm isotropic resolution following the functional scans collected during task play.

fMRI data preprocessing and analysis

Data was preprocessed using a standard pipeline for preprocessing of multiband data. Using FSL (Smith et al., 2004), images were brain extracted, and denoised using ICA component removal, where components were extracted using FSLs Melodic, and classified into signal or noise with a classifier trained on independent datasets. Functional data was then aligned, high-pass filtered (100 s threshold), and unwarped. T2 images were aligned to T1 images with FSL FLIRT, then both were normalized to standard space using ANTs (using CIT168 high resolution T1 and T2 templates (Avants et al., 2009; Tyszka and Pauli, 2016)). Functional data was co-registered to anatomical images using FSLs FLIRT, then registered to the normalized T2 using ANTs. Finally, the functional data was smoothed using a 8 mm FWHM Gaussian kernel. GLMs were specified using default specifications in SPM 12 (Penny et al., 2011). The details of each first level GLM are provided in the main text. Second level T-maps were constructed by combining each subjects first level contrasts with the standard summary statistics approach to random-effects analysis implemented in SPM. Statistic images were assessed for cluster-wise significance using a voxel-height threshold of $p < 0.001$ and SPM's Gaussian Random Field method for brain-volume cluster extent threshold of FWE < 0.05 to identify clusters larger than expected by chance alone.

GLM design for fMRI analysis

The general linear model (GLM) was used to generate voxelwise statistical parametric maps (SPMs) from the fMRI data. The GLMs depicted in Figure 4 included event onset regressors of zero duration (fixation, stimulus, response, and feedback), as well as a boxcar regressor lasting the trials duration. GLM 1 was defined using subject-specific stimulus locked parametric regressors consisting of option utility, stimulus novelty, and estimation uncertainty for both the selected and rejected option, as well as the model estimated reward prediction error as a feedback locked parametric regressor. GLM 1 expressed mean regressor correlations across participants of $|r| < 0.32$ with a mean correlation of $|r| = 0.18$ over all variables (see Figure S3A). The strongest regressor correlation, between selected stimulus utility and selected stimulus uncertainty, ranges between [−0.59, 0.05] across subjects. All significant second-level analysis clusters are reported in Figure S4A, and mean parameter estimates from each cluster reported in Figures 4A–4C are illustrated in Figure S4C.

GLM 2 decomposed each option's utility into its constituent parts, and is defined using subject specific stimulus locked parametric regressors consisting of expected reward value, uncertainty bias, stimulus novelty, and estimation uncertainty for both the selected and rejected option, as well as the model estimated reward prediction error as a feedback locked parametric regressor. GLM 2 expressed mean correlations among regressor across participants of $|r| < 0.59$, with a mean over all variables of $|r| = 0.17$ (see Figure S4B). Of note, we observe a higher correlation between the uncertainty bias and stimulus novelty (min/max subject-level variable correlations: [0.19–0.81]), which is expected since the uncertainty bias term is defined in terms of stimulus novelty and stimulus uncertainty. However, all regressors are entered simultaneously (i.e they are not orthogonalized), meaning all reported effects are significant above and beyond the shared variance. All significant second-level analysis clusters are reported in Figure S4B, and mean parameter estimates from each cluster reported in Figures 4A–4C are illustrated in Figure S4C.

There is a notable correlation between the uncertainty bias and novelty regressors, which is expected given that stimulus novelty plays a multiplicative role in the bias term's value. The GLMs are run without orthogonalization, meaning that variables are forced to compete for variance and the observe uncertainty bias effects are driven by variance unique to that variable. However, the relationship between these predictors could potentially result in an unstable relationship with BOLD activation. To address this concern, we ran a third GLM that did not include the novelty regressors. This analysis reveals a weaker effect of uncertainty bias in mPFC, but critically, the effects remain positive (see Figure S3C). This demonstrates that the observed effects of uncertainty bias in mPFC are not a bi-product of collinearity between stimulus novelty and the uncertainty bias.

Each fMRI session was entered separately for each subjects first level analysis (fixed effects), and regressors were entered without orthogonalization (i.e. simultaneously compete for variance). First-level maps were entered into a second level analyses for the contrasts of interest (random effects) using SPM. Correlations among regressors are illustrated in Figure S3.

Neural correlates of expected reward associated with the linear bonus model

We derive estimated time-courses for variables of interest by fitting the fmUCB model defined to use the linear novelty bonus mechanisms (see Equation 19) to participant behavior. We then applied model estimates as parametric regressors in the GLM outlined in Figure 4 to identify regions associated with stimulus reward valuation (selected + rejected q-value). This analysis identified a cluster in vmPFC that encompassed the cluster associated with q-values estimated by the fmUCB model (see Figure S5A). Regions identified by both models are subsumed by the vmPFC ROI independently identified by meta-analyses (Clithero and Rangel, 2014; Bartra et al., 2013). An analysis of regions correlated with the reward prediction errors generated by both models also revealed largely overlapping correlates in ventral striatum (see Figure S5B).

Neural correlates of the uncertainty bias absent familiarity modulation

The GLM described in Figure 4 was modified to include the uncertainty bias term estimated by the fmUCB model absent familiarity modulation to identify regions associated with value unmodulated uncertainty valuation (selected + rejected uncertainty bias):

$$U_B(s_i) = \omega_t^U \cdot \sigma^2(s_i)_t \qquad \text{(Equation 1)}$$

No significant clusters were found using conventional voxel height $p < 0.001$, but we report a cluster at a liberal threshold of $p < 0.01$. Notably, this cluster overlaps with the region associated with the familiarity modulated uncertainty bias reported in Figure 4B.

Targeted analyses reveal robust novelty disruption of uncertainty bias processes

Using Bayesian model selection, Figure 6A shows that the familiarity modulated uncertainty bias used by the fmUCB model offers a better explanation of neural data in the ROI than did a model using a non-modulated uncertainty bias. However, since the uncertainty bias term used by each model differed only by the presence or absence of the multiplicative novelty term, both variables were highly corrected when all task trials were considered for analysis ($r = 0.8$). Noting that the two uncertainty bias terms differ with respect to the multiplicative novelty term, we identify trials in which a novel stimulus was offered (defined as 3 or fewer observations), as critical trials of differentiation. Limiting our analysis to these trials reduced the correlation between the two uncertainty bias terms to $r = 0.3$.

We conducted two additional Bayesian model comparisons in which we restricted the analyses to these critical trials to demonstrate that the reported results do not spuriously favor the fmUCB model due to lack of identifiability. Using a 10 mm spherical ROI centralized over the uncertainty bias cluster reported in Figure 4B (x = 0, y = 42, z = 0), the first model included only the uncertainty bias term for both the selected and rejected options as a parametric modulator associated with the stimulus onset regressor, thus avoiding potential inconsistencies in shared variance with other variables used by the computational models. Using an exceedance probability threshold > 0.9, and an extent threshold of 10 voxels, 80% of voxels (199 of 246) favored the fmUCB's familiarity modulated uncertainty bias term. No voxels survived this threshold in favor of the non-modulated uncertainty bias term, though 2 voxels (1%) matched the exceedance probability threshold = 0.9 absent extent threshold.

We conducted a second model comparison in the same ROI focusing on variance unique to the uncertainty bias terms by including additional variables used by the computational model (expected reward value, stimulus novelty, stimulus uncertainty) as parametric modulators associated with stimulus onset. Using an exceedance probability threshold > 0.9, and an extent threshold of 10 voxels, 97% of voxels (239 of 246) favored the fmUCB's familiarity modulated uncertainty bias term. No voxels survived this threshold in favor of the non-modulated uncertainty bias term regardless of extent threshold.

In summary, focusing the analysis on the set of trials that maximally differentiate the two mechanisms, we observe stronger indications that the uncertainty bias term in mPFC is blunted by novelty.

Regions of interest and small volume correction

The 15 mm radius sphere in vmPFC used for analyses reported in Figure 5B centered on peak vmPFC voxel reported in a meta analysis of the neural correlates of decision making and subjective utility [x = −2, y = 40, z = −6] (Clithero and Rangel, 2014). Center coordinates for left ([x = −14, y = 10, z = −6]) and right ([x = 14, y = 10, z = −6]) ventral striatum were defined using coordinates from NeuroSynth, while the ROI was defined as the union of 15 mm spheres centered at both left and right coordinates in union with voxels labeled as comprising putamen or accumbens according to the Harvard-Oxford sub-cortical structural atlas. Spherical small volume corrected analysis applied to novelty-biased reward prediction errors reported in Figure 5A used a 9 mm sphere centered on coordinates [x = 18, y = 16, z = −10] as previously reported by Wittmann et al. (2008).

FPC ROIs were defined using 5 mm radius spheres centered on peak voxels in left and right FPC as reported by Daw et al. (2006). The analysis reported in Figures 6B and 6C illustrates the activation patterns in FPC as a function of exploratory choice targeting value, uncertainty and novelty at the time a response is made. These effects are also present if the analysis is locked to the time of stimulus presentation. Using the same ROIs and the same model design using stimulus-locked variables reveals significant positive activation associated with exploration targeting lower valued options ($t(31) = 2.57, p = 0.01$) and more uncertain options ($t(31) = 4.06, p = 0.0003$), but significantly reduced activation when more novel options where chosen ($t(31) = -3.10, p = 0.004$). Our experimental design did not attempt to explicitly dissociated the neural time-course associated with decision and response, and as such, it remains unclear as to whether FPC is influencing the decision making process or engaging a specific action strategy.

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavioral analysis

Behavioral data was analyzed using mixed-effects logistic regression for a descriptive characterization of task performance (using lme4 package in R). We define each option's expected value ($E[S_i]$) as the mean of a Beta distribution specified according to the number of wins and losses observed within the current block of trials (Beta[α = number of wins +1; β = number of losses +1]). Uncertainty $(U[S_i])$ was defined as the variance of the same Beta distribution. We define stimulus novelty $(N[S_i])$ as the variance of a Beta distribution specified according to the number of times a particular stimulus had been observed across the entire experiment (Beta[α = number of exposures +1, β = 1]).

Probability of selecting the option presented on the left $(a_t = L)$ for each trial t was modeled as:

$$p(a_t = L) = (R_\Delta + U_\Delta + N_\Delta) * t + ((R_\Delta + U_\Delta + N_\Delta) * t|ID) \qquad \text{(Equation 2)}$$

where $R_\Delta = E[S_L] - E[S_R]$ denotes the reward differential in favor of the left option, while $U_\Delta = U[S_L] - U[S_R]$, and $N_\Delta = N[S_L] - N[S_R]$ reflect the difference in uncertainty and novelty respectively. We include unit scaled trial number t [ranging from 0-1] as an interaction term to model changes in feature influence across the block horizon, with random intercept and slopes estimated all terms for each participant ID. Note that this configuration implies that main effects of R, U, and N reflect the estimated effect at the start of each block when $t = 0$.

The regression analysis depicted in Figure 2D reports a decreasing effect of expected value across trials. This effect does not necessarily reflect a behavioral tendency to exploit learned values less as the learning context progresses; rather, it is consistent with value integration that deviates from optimal Bayesian integration. To demonstrate this, we used the fmUCB model to generate choices using three different learning rates (100 simulated participants at $\eta = [0, 0.1, 0.25]$). We then applied the regression model to data at each value of η and report the predicted effect of expected value at 10 points across a block of trials (see Figure S1B). When $\eta = 0$ and the fmUCB model optimally integrates observed outcomes, the regression reports a consistent effect of expected value across trials, reflecting agreement between the expected values used by the computational and regression models. However, as the computational model grows increasingly forgetful ($\eta \rightarrow 1$), the regression reports an increasingly negative EV:t interaction, illustrated here as a decreasing effect of EV across trials (blue and green). This decreasing EV:t interaction term reflects a growing disagreement between expected values derived by the forgetful computational model and the optimal regression model, resulting in a decreasing effect of EV.

Analysis of task comprehension

Our experimental design decoupled estimation uncertainty and stimulus novelty by framing each block of trials as a unique learning context, and as such, the value associated with a stimulus sampled during previous contexts should not be applied to the current context in play. Participants were explicitly informed of this, and told that although they may see the same slot machines in different casinos, each casino has programmed the winning probabilities for each machine differently. Here we demonstrate that participant behavior faithfully reflects *de novo* learning in each context, and thus our assumption that the expected value and uncertainty associated with stimuli encountered during previous contexts can be appropriately modeled using an unbiased prior at the start of a learning block (Beta[$\alpha = \beta = 1$]).

We probed for behavioral signatures of value carry-over from previous contexts by augmenting the fmUCB model to accommodate value carry-over from previously learned stimulus values. To do so, we define the Beta distribution describing stimulus i's expected value according to:

$$\alpha_i^* = 1 + \sum_{t=0}^{T-1} \left(\eta^{T-t} \cdot O_t^w\right) + \left(\eta^T * \omega_{prev} * \alpha_{i:prev}^*\right) \qquad \text{(Equation 3)}$$

$$\beta_i^* = 1 + \sum_{t=0}^{T-1} \eta^{T-t} \cdot O_t^L + \left(\eta^T * \omega_{prev} * \beta_{i:prev}^*\right) \qquad \text{(Equation 4)}$$

Here, the expected value is derived according to wins and losses observed in the current context up to the current trial T, where η is a free parameter regulating the rate at which previous outcomes observed in the current context are down-weighted in favor of more recent observations (just as the 'forgetting rate' defined in the fmUCB model). The Beta distribution's parameters for stimulus i also include some proportion of the value associated with that stimulus in the most recent previous context ($\alpha_{i:prev}^*$ and $\beta_{i:prev}^*$), where ω_{prev} governs the proportion of value carry over, and η^T down-weights the initialization value as trials proceed in order to accommodate temporal effects of learning.

We fit and submitted the fmUCB model, a model with $\omega_{prev} = 1$, and a model with ω_{prev} as a free parameter to a model comparison using the CBM toolkit. This comparison showed that the original fmUBC, absent any mechanism to carry learned values across contexts, fit the data best (exceedance probability = 0.97), showing that participants did indeed adhere to the instructed task structure.

This comparison across computational models offers evidence that participants did not carry values across contexts as they were instructed to do. However, we wanted to ensure that the influence of previous context wasn't simply misattriuted and accommodated by other mechanisms in the model. To address these concerns we implemented a sliding window regression analysis, conducting a model comparison between a model that included values from previously encountered contexts to a model that didn't. We defined a baseline model as:

$$p(a_t = L) = R_\Delta + U_\Delta + N_\Delta + (1 + R_\Delta + U_\Delta + N_\Delta|ID) \qquad \text{(Equation 5)}$$

where $R_\Delta = E[S_L] - E[S_R]$ denotes the reward differential in favor of the left option, while $U_\Delta = U[S_L] - U[S_R]$, and $N_\Delta = N[S_L] - N[S_R]$ reflect the difference in uncertainty and novelty respectively, with random intercept and slopes estimated for each participant ID, with values defined as they were for Equation 2. We then defined an augmented model that also included the expected value leaned from the most recent context:

$$p(a_t = L) = PrevR_\Delta + R_\Delta + U_\Delta + N_\Delta + (1 + PrevR_\Delta + R_\Delta + U_\Delta + N_\Delta | ID) \quad \text{(Equation 6)}$$

where $PrevR_\Delta$ denotes the previous context's expected value differential between left and right stimuli. We then repeatedly fit and compared the variance explained by both models across a sliding window of two trials within blocks (e.g. window 1 = trials 1 and 2, window 2 = trials 2 and 3, etc…). This analysis showed that the previous value did not offer sufficient additional explanatory power for any window of trials (all p values > 0.05, uncorrected for multiple comparisons).

In summary, neither computational model comparison, nor computationally agnostic analysis of fine grained structure embedded in participant's game-play showed evidence of value carry-over from previous contexts, demonstrating that participant behavior reflected the instructed task structure faithfully.

Analysis of nuisance task variables

Participants exhibited a robust preference for novel options that increased as they progressed through a learning context. The fmUCB model reproduces this phenomena by combining a constant novelty bias pulling participants toward novel options with a growing push away from paired uncertain familiar options. However, colloquial interpretations such as growing boredom, or a superstition that that novel options might offer a bonus may also tempt inquiry.

Participants may have adopted the unfounded belief that novel options were baited to offer a bonus. Participants were not offered any instructions to hint at this, nor was there an empirical difference in the rewards experienced after sampling a novel or familiar option for the first time (mean reward for both familiar and novel options = 0.5, $t(31) = -0.2, p = 0.85$). Furthermore, given that participants exhibited a growing preference for novel options, they would presumably need to assume that novel options presented later in a block of trials was more likely to offer a baited bonus than novel options presented earlier. No participants reported such a strategy or belief.

Participants were in the scanner for approximately 75 min while functional and structural scans were collected, with a total time on task of approximately 60 min and an average time of just under 4 min per learning block. Most participants reported finding the task challenging and relatively fun. However, to demonstrate that behavior wasn't influence by fatigue, boredom, or other anomalous time-on-task phenomena, we probed for the emergence of shifting strategies as the experimental task progressed. We modified the regression model defined in Equation 2 to include block number as an interaction term instead of trial number.

$$p(a_t = L) = (R_\Delta + U_\Delta + N_\Delta) * b + ((R_\Delta + U_\Delta + N_\Delta) * b | ID) \quad \text{(Equation 7)}$$

Model comparison showed that trial number offered a significantly better predictor of choice than block number (log-likelihoods −6515.5 and −6557.9 respectively), demonstrating that preferences for novelty, uncertainty or rewarded stimuli did not shift meaningfully as the experiment progressed. We constructed a second comparative model by augmenting the model described by Equation 2 to include an additional term specifically probing for an effect of block on novelty seeking:

$$p(a_t = L) = (R_\Delta + U_\Delta + N_\Delta) * t + (N_\Delta * b) + ((R_\Delta + U_\Delta + N_\Delta) * t + (N_\Delta * b) | ID) \quad \text{(Equation 8)}$$

Replicating effects previously reported, this model identified a significant novelty seeking bias ($\beta_N = 0.18, p < 0.01$) that increased within the block of trials ($\beta_{N:t} = 0.42, p < 0.01$). However, there was no effect of block number ($\beta_b = -0.08, p = 0.44$) nor was there a significant interaction with novelty ($\beta_{N:b} = -0.05, p = 0.6$). Furthermore, model comparison showed that the additional block number variable was unwarranted ($\chi^2(21, 32) = 20.528, p = 0.49$), demonstrating that novelty seeking strategies are not accounted for by experiment duration (as opposed to block level) variables.

Computational modeling of behavior

We characterize the computational mechanisms balancing the trade-off between exploration and exploitation using a forgetful Bayesian model of choice, augmented with an uncertainty bias and optimistic value initialization. The subjective utility derived for each stimulus s_i is defined as:

$$V(s_i) = Q_N(s_i) + U_B(s_i) \quad \text{(Equation 9)}$$

where $U_B(s_i)$ is the uncertainty bias added to the optimistically initialized expected value, $Q_N(s_i)$. The probability of selecting either the option presented on the left (s_l) or right (s_r) was derived using a Softmax function, meaning choice was a function of both random and uncertainty-directed exploration:

$$p(s_l) \propto \beta * (V(s_l) - V(s_r)) \quad \text{(Equation 10)}$$

where β is the Softmax parameter controlling the degree to which choice was determined by value.

The forgetful Bayesian reinforcement learning agent maintains a representation of each slot machine in a given block of trials as a Beta distribution. This distribution was defined according to a recency weighted integration of observed outcomes:

$$\alpha_i^* = 1 + \sum_{t=0}^{T-1} \eta^{T-t} \cdot O_t^w \qquad \text{(Equation 11)}$$

$$\beta_i^* = 1 + \sum_{t=0}^{T-1} \eta^{T-t} \cdot O_t^L \qquad \text{(Equation 12)}$$

where T denotes the current trial within the block, and O_t^w and O_t^L are binary flags noting whether or not the observed outcome on trial t was a win or loss respectively. Thus, η operates as a forgetting rate, controlling the rate at which past outcomes are down-weighted in favor of more recent outcomes. Each option's expected value was derived as the mean of this distribution:

$$Q(s_i)_t = \frac{\alpha_{s_i}^*}{\alpha_{s_i}^* + \beta_{s_i}^*} \qquad \text{(Equation 13)}$$

while uncertainty was derived as the variance of the same distribution:

$$\sigma^2(s_i)_t = \frac{\alpha_{s_i}^* \cdot \beta_{s_i}^*}{\left(\alpha_{s_i}^* + \beta_{s_i}^*\right)^2 \cdot \left(\alpha_{s_i}^* + \beta_{a_i}^* + 1\right)} \qquad \text{(Equation 14)}$$

Optimistic initialization was integrated into the model by way of inflating the hyper-parameters describing the expected value ($\alpha_{t=0}^* > 1$ for a novelty seeking bias, and $\beta_{t=0}^* > 1$ for a novelty avoidance bias). For example, a positive novelty bias is embodied by adding a bonus initialization value to $\alpha > 1$, which is correspondingly decayed across trials according to the decay rate defined by η. As such, the boosted value of α endows the distribution representing the novel option with an elevated expected reward value, and a corresponding reduction in uncertainty prior to observing an outcome.

Both the novelty-induced optimistic initialization bias and the uncertainty bias were subject to dynamic weighting terms defined according to the block's horizon. On each trial, weighting terms ω_t^σ and ω_t^N are applied to the uncertainty bias and optimistic initialization values respectively, where weights are defined as a linear function of the current block's trial number:

$$\omega_t^U = \frac{U_I + t \cdot (U_T - U_I)}{T_{task}} \qquad \text{(Equation 15)}$$

$$\omega_t^N = \frac{N_I + t \cdot (N_T - N_I)}{T_{task}} \qquad \text{(Equation 16)}$$

where T_{task} denotes the maximum task horizon, U_I and N_I denote intercepts for uncertainty and novelty at the start of each block, while U_T and N_T denote weights at the end of each block. Thus, ω_t^U and ω_t^N reflect linear trajectories across the task horizon.

We define the fmUCB model as a resolution to the growing tension between novelty seeking and uncertainty aversion. We embody this mechanisms by way of a familiarity modulated uncertainty bias. Stimulus familiarity was defined according to the normalized variance of a Beta distribution defined according to hyper parameters (Beta[$\alpha(s_i)$ = number of observations +1, and $\beta = 1$]), or specifically:

$$F(s_i) = 1 - \frac{\alpha(s_i) \cdot \beta}{(\alpha(s_i) + \beta)^2 \cdot (\alpha(s_i) + \beta + 1)} \qquad \text{(Equation 17)}$$

in which the variance term is scaled to range between [0,1]. We then augment the uncertainty bias to also reflect stimulus familiarity:

$$U_B^f(s_i) = F(s_i) \cdot \left(\omega_t^U \cdot \sigma^2(s_i)_t\right) \qquad \text{(Equation 18)}$$

Lastly, in contrast to novelty directly impacting the expected value, we also test a mechanism in which novelty is factored into the decision making process as its own bonus feature, referred to as the linear bonus model. In this model the optimistic initialization mechanism was removed from the fmUCB model, and augment the subjective utility to include a novelty bonus term:

$$V(s_i) = Q(s_I) + U_B^f(s_i) + \omega_t^N \cdot (1 - F(s_i)) \qquad \text{(Equation 19)}$$

Parameter estimation and model comparison

All model parameters were estimated using the Computational Behavioral Modeling (CBM) toolkit, a hierarchical Bayesian inference method to support model fitting and model comparison within the same framework (Piray et al., 2019). Parameter estimation and model comparison proceeded by first fitting each model of interest to each subject separately (i.e non-hierarchically). In a second hierarchical step, empirical priors are constructed from the first-level fits and mean-field variational Bayes expectation maximization is iteratively applied in which 1) summary statistics are calculated, 2) group parameter posterior estimates are updated, 3) individual parameter posterior estimate are updated, and 4) the responsibility of each model in generating individual data is updated.

Importantly, this applies a random effects approach to model identification, which allows the fitting processes to accommodate the fact that different models might underlie data in different subjects. This has implications for both model comparison and for parameter estimation, as the contribution of a specific model to the group parameters is weighted according to the responsibility of that model generating data in a given subject.

The toolkit relies on normally distributed parameters, meaning some model parameters need to be transformed to be sensible. Novelty and uncertainty weighting terms (N_I, N_T, U_I, U_T) all remained normally distributed (no transformation). The Softmax inverse temperate was constrained to range between $[0 \leq \beta \leq 20]$, while the learning rate was constrained to range between $[0 \leq \ \eta \leq \ 1]$ using a sigmoid function.

Model comparison was also relied on the CBM toolkit (Piray et al., 2019). First-level estimates were computed for each individual and model using common priors $(\mathcal{N}(\mu \ = \ 0, \sigma^2 \ = \ 6.25))$, which were then used to inform second-level fits and simultaneous model comparison. This process treats model comparison as a random effect (i.e. different models might better represent different participant), while also taking advantage of hierarchical parameter estimation which relies on empirical as opposed to prescribed priors.

We conducted a parameter identifiability analysis by using the fmUCB model to generate data using a known set of parameters, then fitting the fmUCB model to that generated data to derive parameter estimates (see Figure S2A). Estimated parameters were highly correlated with the true generative parameters, with all $r > 0.85$. The novelty bias term (N_I) had the lowest correlation ($r \ = \ 0.85$), with all other variables having a correlation of $r > 0.95$. The lower correlation for N_I was associated primarily with high forgetting rates $(\eta > 0.5)$, where the novelty initialization bias is quickly eroded. However it is important to note that η can be reliably identified, and was typically well below 0.5 when estimated from the behavioral data.

We also conducted a model identifiability analysis across the mechanisms of interest (i.e optimistic initialization and the additive uncertainty bias) to ensure that the model responsible for generating the data could be reliably identified. To do so, we generated data from 5 model configurations that embody and contrast the key computational mechanisms; 1) a baseline model consisting of β and η parameters; 2) a novelty bias model that augmented the baseline model to also include N_I and N_T; 3) an uncertainty bias model that augmented the baseline model to also include U_I and U_T; 4) a novelty + uncertainty bias model that augmented the baseline model to also include N_I, N_T, U_I and U_T; and 5) the fmUCB model which augmented the baseline model to include N_I, U_I, U_T, and the familiarity gating mechanism. We generated 50 simulated participants from each of the five model configurations, sampling parameters from a uniform distributions of $1 < \beta < 15$, $0 < \eta < 1$, $-\ 1 < N_I < 1$, $-\ 1 < N_T < 1$, $-\ \ 1 < U_I < 1$, $-\ \ 1 < U_T < 1$. We then fit the same five model configurations to the generated data using the CBM toolkit, and extracted the log-evidence (which takes the number of free parameters into account) for each model.

As illustrated in Figure S2B, each of the five generative models were correctly identified. For example, when data was generated from the novelty + uncertainty bias model, the same model was found to have the lowest error according to the log-evidence in support of each candidate model. Importantly, this analysis shows that model flexibility (i.e. the number of free parameters) is appropriately accounted for, with the simplest baseline generative processes being correctly identified. Furthermore, we see that novelty and uncertainty bias mechanisms can be uniquely identified, with each respective model being correctly identified, in addition to the model that includes both mechanisms simultaneously. Finally, we see that the unique signature of the fmUCB model's uncertainty bias is correctly identified over other model variants.

We considered additional computational mechanisms in conjunction with and as alternatives to the optimistic initialization and uncertainty bias in order to rule out alternative explanations of the behavioral and neural data. This includes choice 'stickiness', both at the action (left vs. right) and stimulus levels of response, in which autocorrelations among choices emerge independent of reward history (Lau and Glimcher, 2005). We also considered a model that included independent learning rates for win/no-win trials, instantiated as unique 'forgetting' rate parameters (η_G and η_L). Lastly, we compared more traditional reinforcement learning (i.e. non-Bayesian) that maintained independent representations of value and uncertainty. None of which fit the data better than the fmUCB model (all exceedance probabilities $\approx$ 0). We also note that fmUCB model parameter estimates remained largely unaffected when the model was augmented to include sticky choice mechanisms, and the time-courses of computational variables of interest (e.g. the chosen option's uncertainty bias) between augmented and fmUCB models were indistinguishable (mean correlation $r > \ 0.98$).

Behavioral confusability analysis

Discriminability between the optimistic initialization and the policy bias mechanism was probed via model confusability. We first fit the fmUCB model defined in terms of both the optimistic initialization and the policy bias mechanisms to participant data, and used those optimized parameter estimates to specify each model's free parameters. Each model instantiation was then exposed to the same set of trials experienced by our participants, and generated a choice on each trial. This process was repeated 100 times for each participant, resulting in 100 simulated experiments with data generated by each of the two mechanisms. Finally, we fit both implementations to both set of simulated experiments to quantify the proportion of fits that correctly identify the true generative mechanism. This analysis revealed that data generated by the optimistic initialization mechanism was only correctly identified for 53% ± 0.14 of the fits. Data generated by the policy bias mechanism suffered equally poor identification, with 57% ± 1.8 of the fits correctly identifying the generative mechanism. Thus, we conclude that behavior alone cannot distinguish between either the optimistic or policy bias mechanisms given our experimental design.