It's new, but is it good? How generalization and uncertainty guide the exploration of novel options

Hrvoje Stojić*
Universitat Pompeu Fabra

Eric Schulz
Harvard University

Pantelis P. Analytis
University of Southern Denmark

Maarten Speekenbrink
University College London

How do people decide whether to try out novel options as opposed to tried-and-tested ones? We argue that they infer a novel option's reward from contextual information learned from functional relations and take uncertainty into account when making a decision. We propose a Bayesian optimization model to describe their learning and decision making. This model relies on similarity-based learning of functional relationships between features and rewards, and a choice rule that balances exploration and exploitation by combining predicted rewards and the uncertainty of these predictions. Our model makes two main predictions. First, decision makers who learn functional relationships will generalize based on the learned reward function, choosing novel options only if their predicted reward is high. Second, they will take uncertainty about the function into account, and prefer novel options that can reduce this uncertainty. We test these predictions in three preregistered experiments in which we examine participants' preferences for novel options using a feature-based multi-armed bandit task in which rewards are a noisy function of observable features. Our results reveal strong evidence for functional exploration and moderate evidence for uncertainty-guided exploration. However, whether or not participants chose a novel option also depended on their attention, as well as reflecting on the value of the options. These results advance our understanding of people's reactions in the face of novelty.

Keywords: Novelty; Decision making; Reinforcement learning; Function learning; Exploration-exploitation; Open Data; Preregistered Experiments

> Novelty has charms that our minds can hardly withstand.— William Makepeace Thackeray.

As it is late, you are hungry, and your fridge is empty, you decide to go out for dinner. As you make your way towards your favorite restaurant in the area, you notice a new restaurant has just opened down the street. How do you go about choosing between this new option and the tried-and-tested one you have visited so many times before? Our lives are full of choices that involve countless options we have never experienced before. Yet we frequently succeed in trying options that are both novel and good. How do we construct expectations for such novel options? And how do we decide whether or not to try them?

Humans and other animals often display a tendency to explore novel and unfamiliar options. People prefer novel stimuli to predictable ones in a lab setting (Berlyne, 1970), novelty attracts attention in both children and adults (Nunnally & Lemond, 1974) and biases the retrieval of episodes from memory such that higher value is attached to novel episodes (Carpenter & Schacter, 2016). In a consumer setting, people prefer newly-packaged goods over the same goods in old packaging (Steenkamp & Gielens, 2003), and some consumers, so-called early adopters, tend to be the first to try newly-launched products (Mahajan, Muller, & Srivastava, 1990; Rogers, 2010). In animals, rats explore novel environments in the absence of extrinsic motivators (Tolman & Honzik, 1930) and can even withstand electroshocks (Nissen, 1930) or forgo cocaine reward (Reichel & Bevins, 2008) to experience novel options, while monkeys can trade reward for novel information (Blanchard, Hayden, & Bromberg-Martin, 2015).

The tendency to seek out novel options can be beneficial: a novel option's reward is uncertain and may be higher than the reward of familiar options. Thus, exploring novel options can help you make better choices in the future. However, exploration comes with a potential cost. If the option turns out inferior to familiar options, you have foregone the opportunity for higher rewards. This frames the well-known exploration-exploitation dilemma. Should you choose an option that you know and currently like best? Or should you be curious and try a more uncertain option in order to learn about it?

The optimal resolution to this dilemma is tractable only in restricted situations (e.g. through so-called Gittins indices; Gittins, 1979; Whittle, 1980). Heuristic solutions are therefore frequently employed. While not optimal, some heuristic strategies are known to work well. One such heuristic strategy is to assign an “uncertainty bonus” to options. This bonus is like a form of optimism, inflating the expected reward of an option by its uncertainty. This uncertainty bonus encourages exploration of lesser-known options (e.g. Kakade & Dayan, 2002). This account resonates well with empirical findings that novel stimuli activate dopaminergic pathways in humans and other animals (Bunzeck & Düzel, 2006; Schultz, 1998). And while early studies did not produce consistent empirical evidence for an uncertainty bonus in human decision making (e.g., Daw, O’Doherty, Dayan, Seymour, & Dolan, 2006; Payzan-LeNestour & Bossaerts, 2011), recent studies have provided converging evidence in favor of uncertainty-guided exploration (Gershman, 2018; Knox, Otto, Stone, & Love, 2012; Schulz, Wu, Ruggeri, & Meder, 2019; Speekenbrink & Konstantinidis, 2015; R. C. Wilson, Geana, White, Ludvig, & Cohen, 2014).

Contrary to the many findings suggesting humans and animals are keen to seek out novel options, there is also evidence for the opposite behavior – a tendency towards novelty avoidance. One example of this comes from research on wild rats who can go days without food, avoiding to interact with newly-introduced options (Cowan, 1976). The *mere exposure effect* is another example: people can prefer a repeatedly presented object over novel ones (Zajonc, 2001). Similarly, in self-directed learning people tend to choose options with known outcomes (Markant, Settles, & Gureckis, 2016), and in supermarkets consumers are “loyal” to brands, willing to pay a price premium for more familiar products (Ching, Erdem, & Keane, 2013; Keller, 2002).

How can humans and animals be sometimes novelty seeking and sometimes novelty averse? To explain the cooccurrence of both phenomena, Teodorescu and Erev (2014) and Gershman and Niv (2015) proposed that novel options are evaluated in the context of general characteristics of the environment in which they occur. If a novel option is introduced in an environment where options are mostly rewarding, this leads to novelty seeking. If a novel option is introduced in an environment where options are mostly not rewarding, this leads to novelty avoidance. We expand upon this account of seemingly disparate results. Specifically, our account is informed by the observation that options in real-world scenarios tend to come with features beyond their shared environment. Consider the example of encountering a newly-opened restaurant. If you are fortunate, then restaurants in town tend to be of high quality. This might cause you to expect the new restaurant to also be of high quality. However, if upon peeking through the window, you see the restaurant has no customers, dirty tables, and packages of microwave pizza in the kitchen, you will likely avoid it. This is because you have learned from past experience that unpopular and unhygienic establishments which use questionable ingredients tend to provide a disappointing dining experience. On the other hand, if you found the place spotless and bustling with clientele, run by an award-winning chef using only fresh and locally-sourced ingredients, you would presumably not hesitate to try it, even if most restaurants in town tend to be awful.

In rich choice environments, where options come with many features, knowledge about how these features relate to reward can be generalized to novel options. If a novel option has features which are similar to those of highly-rewarding options, the novel option is expected to be highly-rewarding as well. This should lead to novelty seeking. If the novel option is similar to non-rewarding options, it can be expected to be of poor quality. This should lead to novelty avoidance. Some of the features may be shared between options, such as the general context in which the options are found. Other features may be unique to options, allowing discrimination between options that occur in the same context. Feature-based generalization allows one to make predictions about the reward a novel option provides. Feature-based generalization requires agents to learn a function which relates features to rewards. People are known to be adept function learners. The cognitive processes underpinning this ability have been widely-studied, both when the outcome is a continuous variable (usually referred to as function learning – see for example Busemeyer, Byun, Delosh, & McDaniel, 1997; Hammond, 1955; Kalish, Lewandowsky, & Kruschke,

Hrvoje Stojić, Department of Economics and Business, Universitat Pompeu Fabra, Carrer Ramon Trias Fargas 25-27, Barcelona, 08005, Spain; Eric Schulz, Department of Psychology, Harvard University, 52 Oxford Street, Cambridge, MA 02138, United States; Pantelis P. Analytis, Danish Institute of Advanced Studies & Department of Management and Marketing, University of Southern Denmark, Campusvej 55, Odense, 5230, Denmark; Maarten Speekenbrink, Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, United Kingdom.

Early versions of this work were presented at CogSci2017, MathPsych2017 and TEAP2019 conferences, and seminars at University College London, Max Planck Institute for Human Development, Stanford University, California Institute of Technology, Chapman University and University of Southern Denmark. Manuscript was made available as a PsyArXiv preprint `https://psyarxiv.com/p6zev/` (Stojic, Schulz, Analytis, & Speekenbrink, 2018a).

Project data, preregistrations and all related materials are publicly available at the Open Science Framework website: `https://osf.io/c8u9t/` (Stojic, Schulz, Analytis, & Speekenbrink, 2018c).

*Correspondence should be addressed to Hrvoje Stojić, now at University College London (h.stojic@ucl.ac.uk)

2004; Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017; Speekenbrink & Shanks, 2010) and when it is a categorical variable (usually referred to as category learning – see for example Juslin, Jones, Olsson, & Winman, 2003; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Medin & Schaffer, 1978; Nosofsky, 1984; Speekenbrink, Channon, & Shanks, 2008).

Normative considerations as well as empirical evidence suggest that uncertainty may play a crucial role in how people choose among novel and time-honored options. This role can go beyond feature-based generalization. Suppose the newly-opened restaurant has live music. Based on your knowledge of the underlying reward function, you might predict the quality of a meal to be similar to other good restaurants in the area. However, because you have never eaten in a restaurant with live music, you may be more inclined to try it out in order to improve your knowledge of how live music affects your dining experience. In essence, exploring options with features for which your inferences are more uncertain will improve your knowledge of the reward function. Functional knowledge changes the nature of the exploration-exploitation dilemma – exploration can be geared towards reducing uncertainty about the reward function, not just the reward of a specific option. In information-rich environments, such functional uncertainty reduction can have greater impact on long-term rewards. This impact is based on the fact that functional knowledge can be generalized to all options.

To explain the varying reactions towards novelty, we need a theoretical framework that places functional knowledge at its heart, while keeping the focus on uncertainty-guided choices. Put differently, whether or not you should approach a novel option should depend on whether your functional knowledge predicts that the option is good or bad, and whether approaching it helps you to improve your functional knowledge. In our previous work, we have proposed a model that has these characteristics. This model consists of (1) a Bayesian function-learning component which relates features to expected rewards and (2) an uncertainty-guided decision component which balances functional expectations of rewards and the associated uncertainty of the acquired functional knowledge (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojic, 2016; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018, see also Acuna and Schrater (2009) and Borji and Itti (2013) for related earlier work on human structure learning and decision making).

In previous work, we provided evidence for the function-learning component, examining various forms that features can take. We have shown that people's choices are guided by features when they come as option-specific features, either explicitly presented (e.g. a restaurant's rating on a popular review website or how nicely it is decorated; Analytis, Kothiyal, & Katsikopoulos, 2014; Stojic, 2016; Stojic, Analytis, & Speekenbrink, 2015; Wu, Schulz, Garvert, Meder, & Schuck, 2018) or implicitly embedded in the location of options (e.g. located in a neighborhood with good restaurants; Wu, Schulz, Speekenbrink, et al., 2018), as well as when they come shared by all options but potentially influence rewards in option-specific ways (e.g. the weather can affect how you evaluate restaurants with or without a terrace Schulz, Konstantinidis, & Speekenbrink, 2018). We have also consistently found evidence for the uncertainty-guided decision component, i.e. that people use intelligent choice strategies that take into account their uncertainty about predicted rewards (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojic, 2016; Wu, Schulz, Speekenbrink, et al., 2018).

In the present study, we use our modeling approach to derive predictions about the behavior toward novel options suddenly appearing in the choice set, situated in information-rich environments where options have observable features predictive of rewards. According to our model, people will exhibit both *functional generalization*, such that they can distinguish between "bad" and "good" novel options, and *functional uncertainty guidance*, such that they will choose novel options more if their functional knowledge is more uncertain. We compare our account to a model which is insensitive to specific features, but learns about the general context in which all options are encountered. The two models make diverging qualitative predictions about reactions to novel options, which we test in three preregistered experiments using a feature-based multi-armed bandit task. To foreshadow our results, we find strong evidence for functional generalization and moderate evidence for functional uncertainty guidance. We also find that whether or not participants choose the novel option depends on attending to the novel option and on further reflection about options' values.

This work provides a bridge between human function learning and reinforcement learning, which have previously been studied in isolation. We believe that addressing both simultaneously is crucial for advancing knowledge about both topics. Function learning has hitherto been studied in prediction tasks where participants are rewarded for making accurate predictions of a function's output from its inputs (e.g., DeLosh, Busemeyer, & McDaniel, 1997; Speekenbrink & Shanks, 2010; von Helversen & Rieskamp, 2008). To do well in such tasks, participants should learn the function over the whole space of possible inputs. By focusing on function learning in a reinforcement learning context, by contrast, we can discover how people learn functions when this is not the explicit goal, when functional knowledge instead supports determining good actions. Since most options come with observable features, explaining how humans learn feature-reward functions and generalize this functional knowledge to new situations is likely to provide general insights into human experiential decision-making. In realistic situations, knowledge of a function may only need to be accurate in consequential regions, for instance for those feature values which occur often, or as in this study, for feature values which are predictive of high rewards. Because of the exploration-exploitation trade-off – people can only learn about the reward function from those options they choose – this may result in functional knowledge which is purposefully biased towards consequential regions. As resolving the exploration-exploitation dilemma in traditional reinforcement learning settings can lead to predictable biases (Denrell & Le Mens, 2011; Le Mens & Denrell, 2011), we believe that understanding how people resolve the dilemma in settings where options are characterized by features can deepen our understanding of function learning, explain how biased samples are constructed, and thus pave the way for studying the implications of biased sampling in function learning settings for human judgments (see Fiedler, 2000). Moreover, determining how people learn and represent functions can advance our understanding of human reinforcement learning in information-rich environments. Functional knowledge can support the generalization of effective behavior to novel situations. A current focal point in artificial intelligence research is designing algorithms which can usefully transfer learned reward functions to novel tasks (Hassabis, Kumaran, Summerfield, & Botvinick, 2017). Whilst currently difficult for machines, such generalization often appears effortless to humans. Advancing our understanding of human function learning may then not only prove beneficial to understanding how humans learn from their actions in complex tasks, but also to designing effective artificial intelligence.

The feature-based multi-armed bandit task

We study participants' behavior in *feature-based multi-armed bandit tasks*. In a feature-based multi-armed bandit task (FMAB, Stojic, 2016; Stojic et al., 2015), participants are presented with a set of options that each have two observable features and offer an unknown stochastic reward (Figure 1). Participants repeatedly choose between the same options with the goal of accumulating as much reward as possible. The rewards associated to each option depend on the observable features through an initially unknown function. This function can be learned through experience. As in other multi-armed bandit tasks, learning requires participants to trade off between exploration and exploitation. Exploration in our task means choosing options which reduce uncertainty about the function. Exploitation means choosing options that, given current knowledge, are likely to produce high rewards.

Crucially, after 40 trials of choosing between the same nine options, we introduce a novel, tenth option (see Le Mens, Kareev, & Avrahami, 2016, for a similar design). By manipulating the features of the novel option, we can obtain empirical evidence for functional generalization and functional uncertainty guidance in participants' choices.

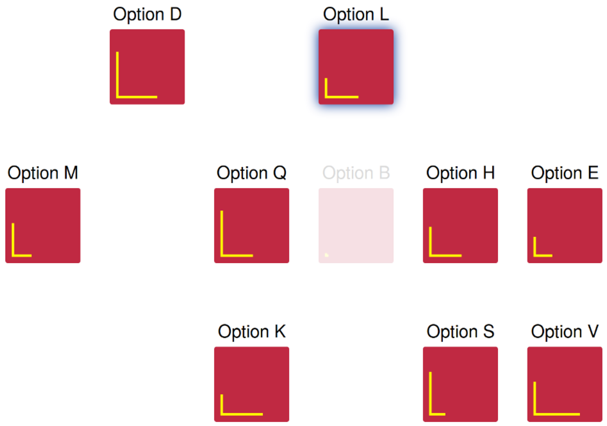


Figure 1. Illustration of a single trial in the FMAB task. Participants choose between options that are presented as red boxes, with the length of horizontal and vertical yellow lines representing feature values. The resulting reward appears immediately below the chosen option. The reward function was a negative linear function of the two features – the smaller the features, the larger the rewards. The same nine options are presented for 40 trials and a novel 10th option appears on the 41st trial in a randomly-chosen position that was previously empty. In the illustration, the novel option is Option B which is transitioning from being transparent to opaque. There are 70 trials in total. In addition to making choices, and before the feedback on the 41st and 70th trial, participants have to estimate the expected rewards of each option and express how confident they are in their estimates.

The general version of this problem is known as a contextual multi-armed bandit problem (e.g. Langford & Zhang, 2008; Li, Chu, Langford, & Schapire, 2010). Scenarios in which the options have option-specific feature values which are predictive of reward through a single function (as in the FMAB problem), and scenarios where the outcomes of different options are influenced by a shared context through option-specific functions (e.g., the location of a town, which affects the quality of seafood restaurants differently than burger joints, as studied by Schulz, Konstantinidis, & Speekenbrink, 2018) are special cases of the general contextual multi-armed bandit framework. A related choice task with multiple dimensions has also been used to study the dynamics of attention in decision making (Niv et al., 2015).

Two strategies for tackling the FMAB task

How will participants react to the introduction of a novel option? Ultimately, their reaction will depend on the strategy they apply. Our functional generalization account assumes that participants will learn how the features relate to the observed rewards. Additionally, their choices will be guided by functional uncertainty – they will balance exploitation with feature-based exploration to reduce their uncertainty about the reward function. We contrast this to a sophisticated reward tracking strategy which –although it ignores feature information altogether– can generalize from experienced rewards across options.

At first glance, it appears irrational to ignore feature information. Yet, people might not be fully aware of the value of generalization, or they might choose this reward tracking strategy because it is less cognitively taxing (Payne, Bettman, & Johnson, 1993). When there are relatively few options and many occasions to choose among them, ignoring the features is relatively harmless. A reward tracking strategy can still learn which options provide high rewards by trying all of them. It can also generalize to novel options by inferring the average reward over all options. However, it cannot distinguish between different novel options. This is because it expects all novel options to have a reward equal to the overall mean.

The difference between how the two strategies generalize also affects how they explore. The reward tracking strategy predicts that all novel options have the same associated uncertainty. By contrast, the functional generalization strategy predicts that the uncertainty about an option's reward depends on the uncertainty about the function at the feature values of that option. The predicted reward for a novel option with feature values similar to already tried options will be less uncertain than for a novel option with dissimilar feature values. While exploration in a reward tracking strategy is geared towards learning the reward of a particular option, exploration in the function-based strategy is geared towards learning the function. The acquired function knowledge enables the function-based strategy to generalize to options with similar feature values.

Functional generalization and uncertainty guidance

Our model of functional generalization and uncertainty guidance combines a flexible Bayesian framework for function learning – Gaussian process (GP) regression (Rasmussen & Williams, 2006; Schulz, Speekenbrink, & Krause, 2018) – with an uncertainty-guided choice strategy – upper confidence bound sampling (UCB, Auer, Cesa-Bianchi, & Fischer, 2002). This model is commonly called GP-UCB (Srinivas, Krause, Kakade, & Seeger, 2012), a name which we adopt here as well.

Gaussian process regression is a Bayesian non-parametric approach towards function learning. It uses a Gaussian process to define a prior distribution over possible functions. It then updates the prior to a posterior distribution over possible functions based on observed inputs (features) and outputs (rewards). Gaussian process regression assumes that outputs y are generated from a function f over (multidimensional) inputs $\mathbf{x}$ and additional noise ϵ:

$$y = f(\mathbf{x}) + \epsilon \quad \text{and} \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon).$$

As a Bayesian technique, prior beliefs about the function f are formalized as a prior distribution over possible functions. The prior distribution is defined as a Gaussian process:

$$f \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right),$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two different inputs. A $\mathcal{GP}$ is parameterized by a mean function $m(\mathbf{x})$:

$$m(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x})\right], \tag{1}$$

which defines the a priori expected value of the output at each input value, and the kernel (or covariance) function $k(\mathbf{x}, \mathbf{x}')$:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right], \tag{2}$$

which defines how the correlation between outputs changes as a function of the difference between the inputs that generated them. The flexibility of GP regression is driven by the kernel function. Choosing a linear or a sinusoidal kernel, functions can be constrained to be linear or periodic. Choosing a radial basis function kernel, functions are allowed to be less regular and more dependent on the particular input values. Different kernels can be thought of as defining different similarity metrics on the inputs. For instance, linear kernels assess inputs as maximally similar when they lie on a straight line. Given a radial basis function kernel, the similarity decreases with Euclidean distance.

As a psychological model of function learning, GP regression incorporates both traditional rule- and exemplar-based accounts of function learning (Lucas, Griffiths, Xu, & Fawcett, 2009). Rule-based accounts assume that people learn functions by assuming the function belongs to a parametric family (e.g., linear, polynomial, or periodic) and then estimating the parameters of the assumed functional family (Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991; Speekenbrink & Shanks, 2010). Exemplar accounts assume that people make functional predictions as a weighted average of previously encountered outputs, where the weights depend on the distance between the input for which a prediction is made and the inputs of the previously encountered outputs (Busemeyer et al., 1997; DeLosh et al., 1997; Kruschke, 1992; Nosofsky, 1986; Speekenbrink & Shanks, 2010). The GP framework incorporates both types of functional representation, either by viewing function learning as a problem of choosing the appropriate kernel – for example, a linear kernel for a rule-based and a radial basis function kernel for a similarity-based account (Lucas et al., 2009), or as finding the appropriate combination of kernels (Schulz et al., 2017).

We use a radial basis function (RBF) kernel to derive a priori predictions. An RBF kernel is defined as

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right), \tag{3}$$

where the signal variance σ_f^2 reflects the average distance of the function away from its mean and the length-scale λ reflects the smoothness of the function (the magnitude of the correlation between the outputs of two nearby inputs). A GP with the RBF kernel has appealing theoretical properties – it is a universal function approximator which is able to learn a wide range of stationary functions (Neal, 1996). Research determining human-like kernels is still ongoing (Lucas, Griffiths, Williams, & Kalish, 2015; Schulz et al., 2017; A. G. Wilson, Dann, Lucas, & Xing, 2015), so we opted for the RBF kernel as a more flexible model of human function learning, closer to exemplar-based learning. Although it is well-known that people are biased towards assuming positive linear functions (Busemeyer et al., 1997; Lucas et al., 2009), they can rely on exemplar-type strategies as well (DeLosh et al., 1997; Juslin, Olsson, & Olsson, 2003). All of our qualitative predictions generalize reasonably well over different choices of kernel function. While effect sizes are affected by using a linear, RBF, or a mixture kernel, the direction of our predicted effects is not. Our choice of an RBF kernel is thus not a strong theoretical commitment.

Based on a set of previously observed input-output pairs, Gaussian process regression infers a posterior distribution over functions. This distribution can be used to predict mean rewards as well as the associated uncertainty in these predictions. Knowing about predictions' uncertainty is crucial to guide exploration. The variance of the posterior distribution over possible functions can be used as a proxy for how much knowledge about the function can be improved by trying an option (Krause, Singh, & Guestrin, 2008). If the function's outputs for a particular input are relatively uncertain, then observing the output for that input will improve predictions not only for that particular input, but also for similar input values. In the current context where the inputs are options defined by feature values and the outputs are the rewards obtained by choosing an option, this maps onto the value of exploring an option.

The upper confidence bound choice rule implements functional uncertainty guidance by adding a multiple of the posterior standard deviation to the posterior mean reward, and choosing the option with the highest resulting value. Let $m_{j,t}$ be the posterior predictive mean for option j at time point t, and $v_{j,t}$ the posterior predictive variance (the posterior predictive mean and variance are the mean and variance of the posterior distribution over possible functions based on all observations up to time $t-1$). The UCB sampling strategy assigns a value or utility $u_{j,t}$ to each option as

$$u_{j,t} = m_{j,t} + \beta\sqrt{v_{j,t}}, \tag{4}$$

e.g., as the sum of the posterior predictive mean reward and a multiple (β) of the uncertainty about the mean reward (the posterior predictive standard deviation). For a normally distributed variable, the second component corresponds to an upper confidence bound; for example, with $\beta = 1.96$, the 95% upper confidence bound.

When the goal is to maximize reward, choosing options with the highest upper confidence bound is intuitive: if the upper confidence bound of one option is larger than that of another, the probability that this option is better than the other may be substantial, even when its posterior predictive mean reward is lower.

As the UCB rule adds a multiple of the uncertainty to each option's mean reward, it is also a formalization of the uncertainty bonus account. This has been termed "directed exploration", to contrast it to "random exploration" (Gershman, 2018; R. C. Wilson et al., 2014). In simple versions of random exploration, options are chosen randomly according to differences in mean rewards or simply a fixed proportion of the time (the so-called softmax and epsilon-greedy methods, Sutton & Barto, 1998). More sophisticated forms of random exploration take uncertainty into account, for example by drawing a random sample from the posterior predictive distribution of mean rewards (Thompson sampling, Thompson, 1933) or the posterior predictive distribution of actual rewards (Speekenbrink & Konstantinidis, 2015). Evidence suggests that both directed and random exploration might work in tandem (Gershman, 2018; Schulz, Wu, et al., 2019; R. C. Wilson et al., 2014). Because we found strong evidence for the UCB rule in previous feature-based multi-armed bandits (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojic, 2016; Wu, Schulz, Speekenbrink, et al., 2018), we focused on the UCB rule to derive our predictions.

To compute $P(C_t = j)$, the probability that the choice C on trial t is option $j \in \{1, \dots, K\}$, we assume a soft maximization:

$$P(C_t = j) = \frac{\exp(u_{j,t})}{\sum_{i=1}^{K} \exp(u_{i,t})}. \tag{5}$$

Note that the stochastic UCB choice rule reduces to a standard Softmax choice rule (with temperature parameter equal to one) if $\beta = 0$. In this case, an option's current predictive uncertainty is not taken into account and exploration essentially happens at random (Sutton & Barto, 1998).

Hierarchical reward generalization and exploration

We contrast the function learning strategy to a reward tracking strategy which ignores the features altogether. The reward tracking strategy assumes that each option is drawn from a common population and treats the rewards associated to each option as otherwise independent from the other options. Following Gershman and Niv (2015), we use a Bayesian hierarchical (BH) model which assumes that people learn about the mean and variance of an option's rewards, while at the same time building up a higher-level representation of the common distribution from which the options were drawn. We again combine the reward tracking learning model with the UCB choice rule, and refer to the resulting model as the BH-UCB model.

As options in our task provide continuous-valued rewards, we use a hierarchical Gaussian model rather than the Bernoulli model put forward by Gershman and Niv (2015). Our hierarchical model assumes that the rewards of each option j are drawn from a Normal distribution

$$R_j^t \sim \mathcal{N}(\mu_j, \sigma_\epsilon), \tag{6}$$

with a common variance σ_ϵ^2 but an option-specific mean μ_j. The option-specific means are assumed to be drawn from a common higher-level Normal distribution

$$\mu_j \sim \mathcal{N}(\mu, \tau), \tag{7}$$

where μ is the average reward over all options, and τ^2 the variance of the option-specific means. The model is completed with prior distributions for μ, σ, and τ, for which we used a $\mathcal{N}(0, 10)$, half-Cauchy(0, 10), and half-Cauchy(0, 10) distribution, respectively (half-Cauchy distributions were truncated below at 0). Having observed rewards of the options, the model updates these to a joint posterior distribution over the means μ_j, the common mean μ, and the variances σ^2 and τ^2. At any time t, the joint posterior distribution provides posterior predictive distributions of the average reward for each option.

Just as for the GP-UCB model, the posterior predictive mean $m_{j,t}$ and variance $v_{j,t}$ are used to compute the UCB values (Equation 4), which are then used to compute choice probabilities using the Softmax function (Equation 5). Given a novel option, the model expects its mean reward to reflect the posterior distribution of μ (i.e., the expected reward for a novel option is thought to be the posterior mean of μ). We implemented the model using RStan (Stan Development Team, 2018).

Participants can perform relatively well in the FMAB task if they employ a reward tracking strategy. In practice, this strategy corresponds to trying out each option a few times and then deciding on the one with the highest expected reward. Because the hierarchical model also infers the distribution from which options' expected rewards are drawn, it is possible to generalize to novel options with a simple rule – novel options are expected to produce a reward that is equal to the mean of the inferred higher-level distribution. Importantly, as the BH-UCB strategy completely ignores the features, it generates the same prediction for any novel option. Gershman and Niv (2015) provided support for a similar model, finding that people indeed generalize their prior experience in a choice environment to make inferences about novel options. This form of experience-based generalization shares characteristics with normalization-based accounts in reinforcement learning, which have been supported by previous research (Louie, Khaw, & Glimcher, 2013; Palminteri, Khamassi, Joffily, & Coricelli, 2015; Rigoli, Friston, & Dolan, 2016). Ignoring the features, while still being able to generalize, makes the BH-UCB model an appropriate competitor to the GP-UCB model.

Experiment 1: Functional generalization

The first preregistered experiment assessed functional generalization by introducing a novel option with features that indicated either low or high rewards. The two strategies, embodied by the GP-UCB and BH-UCB models, will treat these novel options differently. The BH-UCB model is able to generalize in a limited way, by assigning the same expected reward and uncertainty to both novel options. By contrast, GP-UCB is able to distinguish between the novel options and their expected rewards and uncertainty.

We used a between-subject design and a negative linear reward function. In the *FMAB low value* condition, the novel option had high feature values and a resulting low expected reward. In the *FMAB high value* condition, the option had low feature values and a resulting high expected reward.

The experiment had two additional conditions which were equivalent to the FMAB conditions except that the options' features were invisible. This made the task identical to a classic, non-contextual multi-armed bandit (MAB) task. The *MAB low value* and *MAB high value* condition serve as control conditions as they force participants to only learn by using a reward tracking strategy.

Method

Participants. We recruited 320 participants (166 female, M_{age} = 37.1 and SD_{age} = 10.5) through Amazon's Mechanical Turk (`http:\\mturk.com`) online labor market (Crump, McDonnell, & Gureckis, 2013; Paolacci & Chandler, 2014). There were 97 participants in the FMAB high value and exactly as many in FMAB low value condition, 68 in the MAB high value and 58 in the MAB low value condition. We followed a sampling plan based on Bayesian hypothesis testing of our main hypothesis (see Appendix A). Since our main stopping criteria were not met, we stopped collecting the data when we reached the pre-determined budgetary limit. Participants were from the United States and had an approval rate of 95% or higher. We rewarded participants with a fixed payment of $0.70 and a performance-dependent bonus of $1.40 on average. The experiment took 11.9 minutes on average. The study was approved by the UCL Research Ethics Committee.

Feature-based multi-armed bandit (FMAB) task. The task comprised of 70 trials in total. The same 9 options, each characterized by two features, were provided as a choice set until the 41st trial. We refer to these 9 options as the *old options*. On trial $t^* = 41$, an additional option was added to the choice set and thereafter remained available until the end of the task. We refer to the newly introduced option as the *novel option*. We chose to make the novel option appear on the 41st trial to allow for enough time to learn about the underlying function. We chose to let participants sample for 30 more trials after the novel option had appeared (i.e. giving them 70 trial in total), to provide sufficient future opportunity to exploit the new option in case it proved to be good. Furthermore, we settled on an intermediate number of options (i.e. 9+1 options) for which learning functional relations would be feasible, but not overwhelming for the participants who opt for a reward tracking strategy.

Every choice for option k on trial t produced a reward R_k^t associated with that option. Rewards were a negative linear function of an option's features $\mathbf{x}_k = (x_{1,k}, x_{2,k})$:

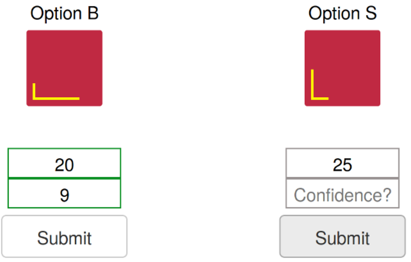
$$\begin{aligned} R_k^t &= f(\mathbf{x}_k) + \epsilon_k^t \\ &= 35 - 20x_{1,k} - 10x_{2,k} + \epsilon_k^t, \end{aligned} \tag{8}$$

where ϵ_k^t was drawn from a Gaussian distribution with a mean of 0 and a variance of 4. We chose a negative linear function to ensure that participants' choices reflect their acquired functional knowledge rather than a potential prior for positive linear functions, which people typically exhibit in function learning experiments (Brehmer, 1974; Busemeyer et al., 1997).

For each participant, the feature values ($x_{1,k}$ and $x_{2,k}$) for the old options were randomly drawn from uniform distributions at the start of the task. These distributions covered three different intervals: $\mathcal{U}(.25, .35)$, $\mathcal{U}(.45, .55)$, and $\mathcal{U}(.65, .75)$, yielding nine possible interval permutations. For example, features for one option were drawn from the $\mathcal{U}(.25, .35)$ and $\mathcal{U}(.25, .35)$ intervals, for another option from the $\mathcal{U}(.25, .35)$ and $\mathcal{U}(.45, .55)$ intervals, and so forth. We randomly sampled feature values to include a wide range of choice sets in our experiment, thus increasing the generalizability of the results. The resulting expected rewards ranged from 12.5 to 27.5. Participants' goal was to maximize the cumulative sum of these rewards during the entire task.

We manipulated (between-subjects) the novel option's features to indicate low or high expected rewards. In the *low value* FMAB and MAB conditions, the novel option had both feature values set to 0.95, resulting in a low expected reward of 6.5 points. In the *high value* conditions, the feature values were both set to 0.05, yielding a high expected reward of 33.5 points.

Estimation task. Our models also generate predictions about options' expected rewards and the associated uncertainty. Examining participants' beliefs about these measures can therefore corroborate the evidence for our predictions derived from the choice data. Hence, in addition to the main task, participants also completed an estimation task on two occasions (Figure 2), where we asked them to estimate the mean reward for each option, as well as rate their confidence in those estimates. We constrained the range for the estimates to be between 0 and 50, while the confidence ratings were entered on a scale from 1 (low confidence) to 10 (high confidence). To ensure that participants provided truthful estimates and meaningful confidence ratings, we rewarded the accuracy of a single estimate at the end of the experiment, where the chance that an estimate was selected was proportional to its confidence rating relative to the other confidence ratings. The earnings depended on accuracy as follows: $\max(0, 300 - 10|E[R_k] - \hat{E}[R_k]|)$, where $E[R_k]$ denotes the true mean reward of option k, and $\hat{E}[R_k]$ the estimate. This reward function was set such that participants could earn a significant amount of money from the two estimation tasks, up to about a third of the total earnings.



20
9

25
Confidence?

Figure 2. Illustration of the estimation task for the FMAB conditions. Participants completed the task on the 41st and the 70th trial, after their choice but before receiving feedback on the reward. They had to estimate the expected reward of each option and express how confident they were in their estimates. The task was identical for the MAB conditions, with the difference that feature values were hidden. Note that the estimate of the mean reward has been entered already in this illustration.

Functional knowledge task. In ongoing research using the FMAB task, we found that a substantial proportion of participants adopt a reward tracking strategy (about 40%, see Stojic, 2016).[1] To distinguish those who learned the function from those who only tracked rewards, participants in the FMAB conditions completed a *functional knowledge task* immediately after the main bandit task. Achieving good performance in this task required participants to generalize functional knowledge they had acquired during the bandit task. Thus, we expected those participants who had learned the function well to achieve better-than-chance performance in this task. Accordingly, our preregistered classification procedure used participants' mean performance in the task to distinguish between *function learners* and *reward trackers*.

The task consisted of 25 trials in which participants had to choose between three options characterized by the same

[1] This is based on experiments where reward functions were also linear, but with different coefficients and feature values. These experiments had similar sample sizes and used clustering methods and statistical tests to classify participants into function learners and reward trackers.

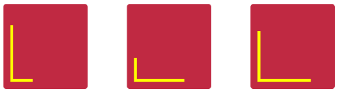


Figure 3. Illustration of a trial in the functional knowledge task. This task follows immediately after the bandit task and only in the FMAB conditions. Features were always visible and rewards were governed by the same function as in the FMAB task. On each trial, participants were asked to choose between three new options and reward feedback was not provided. Options were designed to examine whether participants have learned the reward function – participants with such knowledge should be able to achieve better-than-chance performance.

features as options in the FMAB task (Figure 3). Participants had to choose between three new options on every trial and did not receive feedback about the chosen options' reward. For each of these choice triplets, there was always a best, a medium, and a worst option, assuming perfect knowledge of the underlying function (see Appendix A). We classified participants' performance using a one-tailed Bayesian t-test. Participants who achieved better-than-chance performance were classified as function learners. Participants who did not perform better than chance were classified as reward trackers (see Appendix A).

Given the small number of trials in the functional knowledge task, we expected the classification of function learners and reward trackers to be rather coarse. Thus, the task was designed to be sufficient for distinguishing between the two types of learners, but not for precisely diagnosing participants' knowledge. We performed simulations to confirm that our classification procedure was sufficiently sensitive (see Appendix A). For this, we simulated function learners starting from perfect knowledge, i.e. performing optimally in the functional knowledge task, and made them progressively worse until random performance, i.e. the expected performance level of pure reward trackers. The results of this simulation showed that our classification procedure is sufficient to correctly classify participants with moderate or better functional knowledge as function learners (with 80% probability and higher for mean rank of 1.5 and higher), while those with poorer knowledge are likely to be miss-classified as reward trackers (Figure A1).

Procedure. Participants completed the experiment online. The experiment was programmed in JavaScript and HTML, using the jsPsych (Leeuw, 2015) and Psiturk (Gureckis et al., 2015) libraries. We ensured that participants could only participate once in our experiments by tracking their worker identification numbers.

Participants read the instructions after providing informed consent. We explained that they had to choose between a set of options 70 times, with the goal to earn as many points as possible. We also explained that while the rewards were noisy, the average reward of the options would not change over time. This was done through a metaphor in which each option can be thought of as a bag of coins, and choosing an option means drawing a random coin from the bag. Moreover, we informed participants that there would be additional tasks (the estimation and functional knowledge task) offering another opportunity to increase their earnings in the experiment. Details of these tasks were not further specified in advance. We did not explicitly mention the introduction of the novel option on trial 41. We did stress that the options would remain available once they appeared. We explained in detail how participants' earnings would be determined and that each choice would yield points which were later converted into money at a rate of 1800 points per $1. After reading the instructions, participants completed an attention check questionnaire and were sent back to the instructions if they had answered any of the questions incorrectly (see Appendix B for a brief exploratory analysis of how attention is related to choice performance).

Participants had a maximum of 60 seconds on every trial to select an option. Following each choice, reward feedback was displayed for two seconds, after which the task automatically continued to the next trial.[2] Throughout the task, a counter positioned at the top of the screen displayed the current trial and the total number of trials. In the FMAB conditions, feature values were displayed in the form of a horizontal and a vertical line starting from the lower left corner of the squares representing the options (Figure 1). For example, a feature value of 0.1 would correspond to a short line, while a value of 0.9 would correspond to a line almost spanning the full length of the square. Which line (vertical or horizontal) corresponded to which feature was determined at random for each participant. The features (lines) were not displayed in the MAB conditions. Participants in the FMAB condition were informed that features might be helpful by the following sentence: "Options have horizontal and vertical lines of different lengths drawn inside the squares. The lines can help you predict the value of the coins in each bag.". We did not inform participants about the underlying function. Rather, participants had to infer this function by themselves. Each option had a randomly assigned label to further facilitate their identification. The old options were randomly positioned in a 6-by-3 grid (column-by-row) before the start of the task. The novel option appeared in one of the remaining 9 cells, selected at random, smoothly fading in over a period

[2] The bandit task was supposed to have a one second delay in which participants could not make a response after the trial begins, as outlined in the preregistration. This delay was meant to slow down participants and increase the chance of noticing the novel option appearing. Unfortunately, due to a technical error there was no delay at all, participants could make a response immediately when the trial began.

of three seconds. This was to draw participants' attention to the novel option in case they were looking at other parts of the screen.

In the estimation task, two text input boxes appeared below each option, one to estimate the expected reward and one to rate confidence in the estimate (Figure 2). We presented detailed instructions for this task at the bottom of the screen. Participants completed the task on the 41st and 70th trial, after their choice but before receiving feedback about the reward earned through their choice.

Participants in the FMAB conditions continued with the functional knowledge task immediately after the bandit task (Figure 3). Before the task began, we instructed them that they would have to choose between new options on each trial and that they would not receive reward feedback, but that their final earnings would nonetheless be affected by the reward associated with the chosen options in the same way as in the FMAB task. Each option was placed randomly on a 5-by-1 grid on each trial and the options were unlabeled.

At the end of the experiment, we informed participants about their total earnings, and asked them to report their age, gender, and whether they had noticed that a novel option appeared on the 41st trial (see Appendix B for a brief exploratory analysis of the final question).

Analyses. Detailed overview of statistical analyses can be found in the Appendix A. We use Bayes factors to quantify the relative evidence the data provides in favor of the null (H_0) or the alternative hypothesis (H_1). We denote the Bayes factor that reflects the relative evidence for H_0 compared to H_1 as BF_{01}, and the Bayes factor that reflects the relative evidence for H_1 as BF_{10}. Following (Jeffreys, 1961), we classify a Bayes factor between 3 and 10 as "moderate" evidence in favor of a hypothesis, and a Bayes factor of 10 or larger as "strong" evidence. We mark a test of a preregistered hypothesis with a *-symbol; for example, BF^* indicates the Bayes factor for a preregistered hypothesis. Note that hypotheses that we did not preregister are still often predicted by our model simulations: this will be understandable from the context.

For hypotheses concerning participants' choices of novel options on a single trial we used the contingency table Bayes factor of Jamil et al. (2017), with an independent multinomial sampling assumption and a default "weak" Dirichlet prior ($a = 1$) for H_1. The null hypothesis here was that the allocation of choices does not depend on condition, while the alternative hypothesis was that the choices differ between conditions. To estimate the probability that participants in the FMAB conditions would choose a novel option over the course of multiple trials, we used a Bayesian binomial model: we used a non-centered probit parameterization and our priors of group-level means and standard deviations were determined by our model simulation results (exact prior specifications can be found in the Appendix A).

For hypotheses related to the estimation task and for classifying FMAB participants into function learners and reward trackers, we used a default Bayesian t-test Morey and Rouder (2011); Rouder, Speckman, Sun, Morey, and Iverson (2009), with the Jeffreys–Zellner–Siow prior and scale set to $\sqrt{2}/2$. Since our hypotheses were directional, we used one-sided tests and truncated the prior above or below zero, with the null hypothesis of no difference and the alternative hypothesis of a difference in the predicted direction. We used a symmetric, non-truncated prior whenever we had a non-directional hypothesis, and explicitly indicated when this was the case.

Data and code availability. All project files are publicly available at the Open Science Framework website: `https://osf.io/c8u9t/` (Stojic et al., 2018c). This repository includes the behavioral data, the code used for our model simulations and data analysis, as well as links to all preregistrations.

Predictions

We generated *a priori* predictions by simulating the behavior of both the GP-UCB and the BH-UCB model in our task. We then took the most important patterns, formulated them as hypotheses and preregistered them before data collection commenced (Stojic, Schulz, Analytis, & Speekenbrink, 2018b).

To apply the GP-UCB model to our task, we determined the RBF hyperparameters (σ^2 and λ^2) on each consecutive trial by maximizing the current marginal likelihood. We also subtracted the true mean reward over all options (20 points) to set the prior mean function to 0, simplifying posterior computations. We present simulation results of a UCB choice rule with an exploration parameter of $\beta = 2$ (Figure 4). Simulations for other parameter values ($\beta \in \{0, 1, 3\}$) can be found in the preregistration document (Stojic et al., 2018b).

The simulation results confirm that the GP-UCB model learns the reward function during the first 40 trials and can therefore correctly predict the mean reward of the novel option, resulting in a tendency to choose the novel option when its features indicate high rewards and to ignore it when its features indicate low rewards (Figure 4a, left panels). This behavioral pattern cannot be captured by the BH-UCB model. Furthermore, the GP-UCB model has a higher level of uncertainty regarding the low value novel option than for the high value option. This prediction results from the interaction between function learning and the decision process, because the goal of maximizing rewards biases decision makers to have more experience with options in consequential regions with features associated with high rewards. Consequently, knowledge about good options will be better (more certain) than knowledge about bad ones. After the 41st trial, the GP-UCB model keeps selecting the high value

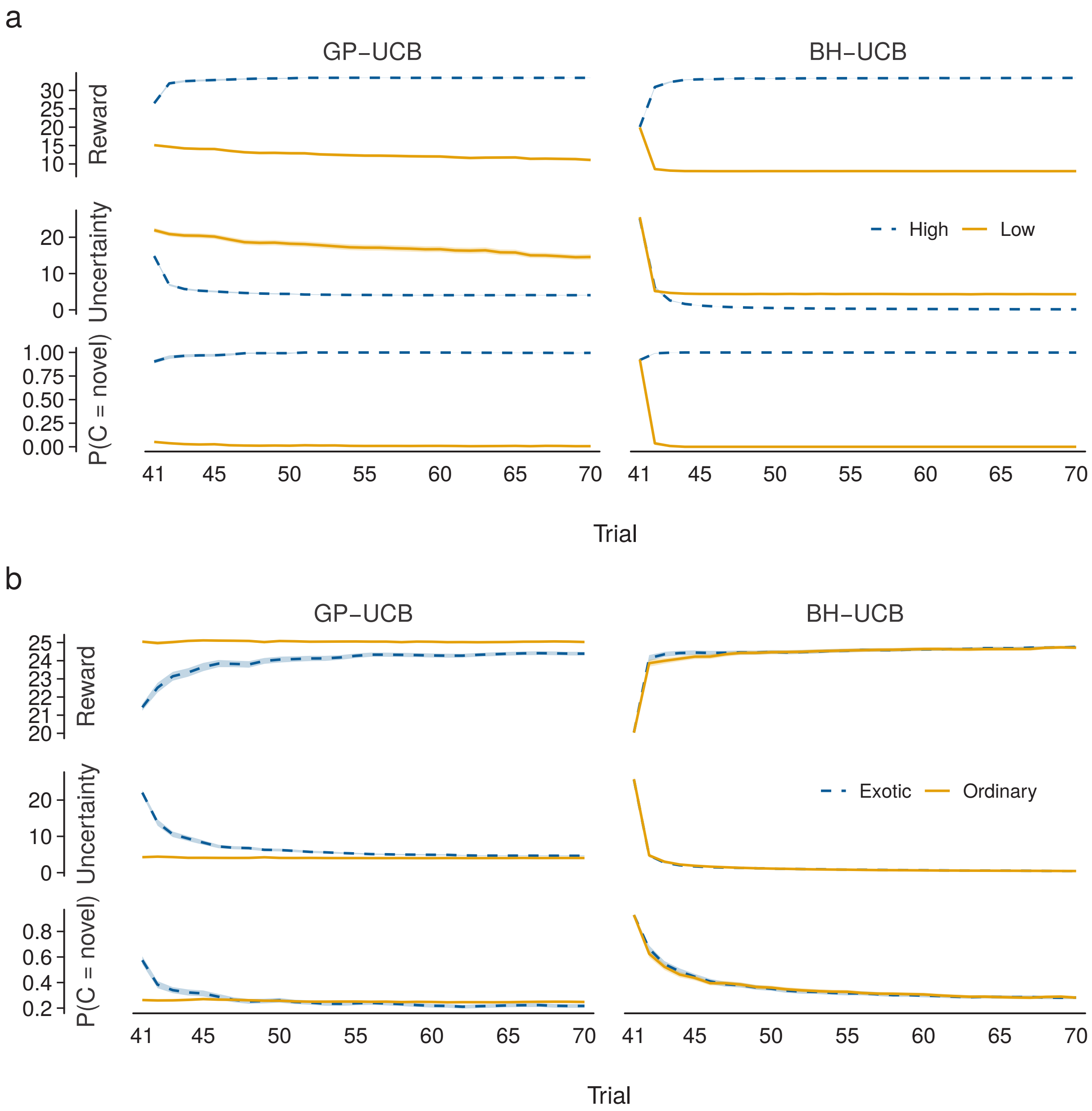


Figure 4. Simulation results for the GP-UCB and BH-UCB model for trials 41 to 70 in Experiment 1 (**a**) and in Experiment 2 (**b**). For each experiment, we show the inferred reward (mean of the posterior predictive distribution, top row) and uncertainty for the novel option (variance of the posterior predictive distribution, middle row), as well as the probability of choosing the novel option (bottom row). Lines represent means across 100 simulations, while bands represent the standard errors of the means. The weight of the uncertainty term was fixed to $\beta = 2$, a medium high value aimed to capture a representative participant.

novel option and its uncertainty reduces, while it ignores the low value option for which uncertainty remains at a high level. The resulting diverging levels of knowledge constitute a novel prediction derived from our framework which does not hold in traditional function learning tasks (Busemeyer et al., 1997; Juslin, Olsson, & Olsson, 2003; Kruschke, 1992; Nosofsky, 1984; Speekenbrink & Shanks, 2010).

Since the BH-UCB model cannot distinguish between novel options before they have been tried, it chooses the high and low value option with the same probability (Figure 4a, right panels) after it is introduced on trial 41. It also assigns the same expected reward and uncertainty to all novel options. If the high value novel option is chosen, it chooses it more frequently thereafter. If the low value novel option is chosen, it chooses it less often thereafter. It is also evident that the BH-UCB model can catch up with the GP-UCB model after a few trials. How rapidly the models converge depends on the magnitude of the exploration parameter, taking longer for smaller values of the exploration parameter (not shown in Figure 4a, but see Stojic et al., 2018b).

The predictions of the GP-UCB model should hold for participants in the FMAB conditions employing a function learning strategy. Participants in the FMAB conditions who ignore the features and employ a reward tracking strategy, as well as participants in the MAB conditions, are expected to behave in line with the predictions from the BH-UCB model. We use the functional knowledge task and our preregistered classification procedure to identify function learners and reward trackers, to be able to examine the model predictions on these more appropriate subgroups of the FMAB conditions.

Results

Choice proportions. One of our primary preregistered hypotheses concerned participants' choices on the 41st trial (Stojic et al., 2018b). Contrary to our hypothesis, participants in the FMAB conditions did not choose the novel option in the high value condition (3%) more frequently than in the low value condition (4%) on trial 41, $BF^{*}_{10} = 0.07$.

Instead of emerging immediately on the 41st trial, the expected difference arose from the 42nd trial onwards (Figure 5a). On the 42nd trial, 30% of FMAB participants chose the high value option and only 6% chose the low value option, $BF_{10} = 2044$. For the MAB participants, those in the high value condition eventually started choosing the novel option more often than participants in the low value condition on trial 47 ($BF_{10} = 10.93$) and onwards (Figure 5d). As predicted by our model simulations, the difference in choice proportion between the high and low value option emerged later than in the FMAB conditions. Moreover, 48.5% of participants did not choose the novel option at all in the FMAB low value condition, while only 21.6% of participants in the high value condition never chose it; a substantial difference with $BF_{10} = 658$.

Based on their performance in the functional knowledge task, 76 out of 194 participants (39.2%) in the FMAB conditions exhibited good knowledge of the function and were classified as function learners. The proportion of function learners was smaller than in our earlier studies using the FMAB task, where it was closer to 60% (Stojic, 2016). Participants' performance in the knowledge task did not cluster around random performance (mean rank equal to two). Instead, their scores spanned the whole range of possible scores. In accordance with our classification simulations, most participants classified as function learners had a good level of performance with a mean rank below 1.5 (Figure B1). The observed variability suggests that performance in the functional knowledge task provided a good basis for classifying participants into function learners and reward trackers (see Appendix B for additional exploratory analyses of our main hypotheses using performance in the functional knowledge task directly rather than classification results).

As our predictions were geared towards function learners, and as there are more reward trackers than function learners in the FMAB conditions, we examined the function learners' behavior separately from the other participants in the FMAB conditions. Focusing solely on function learners, there again was no evidence that the novel high option was chosen more frequently than the novel low option on trial 41 ($BF^{*}_{10} = 0.15$). Instead, the expected difference again arose from the 42nd trial onwards (Figure 5b), with a larger effect than for the FMAB conditions overall – in the 42nd trial 46% of function learners chose the high value option and only 3% chose the low value option ($BF_{10} = 6482$). By contrast, reward trackers were much slower in exploring the high value novel option, with strong evidence for a larger proportion of choices allocated to the high value option (25%) than for the low value option (0%) starting from trial 52 onwards ($BF_{10} = 2978$; Figure 5c), a pattern strikingly similar to that observed in the MAB conditions (Figure 5d).

As an alternative to examining single trials, we also estimated the probability of choosing the novel option on all 30 trials after its introduction (i.e., trial 41 to 70), using a Bayesian binomial model (Figure 5e and Appendix A). The resulting posterior distributions of the probability that participants choose the high and low value novel option showed a clear separation between the FMAB high and low value conditions, with a median of 34.8% (95% credible interval (CI) $[23.2, 47.0]$) for the high value condition and 2.4% (95% CI $[1.6, 3.3]$) for the low value condition. Repeating the same analysis for just function learners yielded an even larger difference, with a median of 64.4% (95% CI $[53.1, 74.7]$) for the high value condition and 1.6% (95% CI $[0.8, 2.5]$) for the low value condition. By contrast, the difference was substantially smaller when comparing the MAB conditions, with a median of 12.9% (95% credible interval (CI) $[5.3, 24.8]$) for the high value condition and 1.5% (95% CI $[0.8, 2.4]$) for the low value condition. We therefore conclude that there is strong evidence that FMAB participants preferred the high value novel option over the low value option.

Expected rewards and uncertainty. Our models also generated differing predictions about options' expected rewards and the associated uncertainty, as assessed in the estimation task on trial 41 and 70.[3] According to the GP-UCB

[3] The estimation task data passed our preregistered sanity checks. We expected that participants' estimated values would be closer to the options' actual expected rewards on trial 70 and that confidence ratings would be higher the more frequently an option was chosen. This was indeed true for both the FMAB

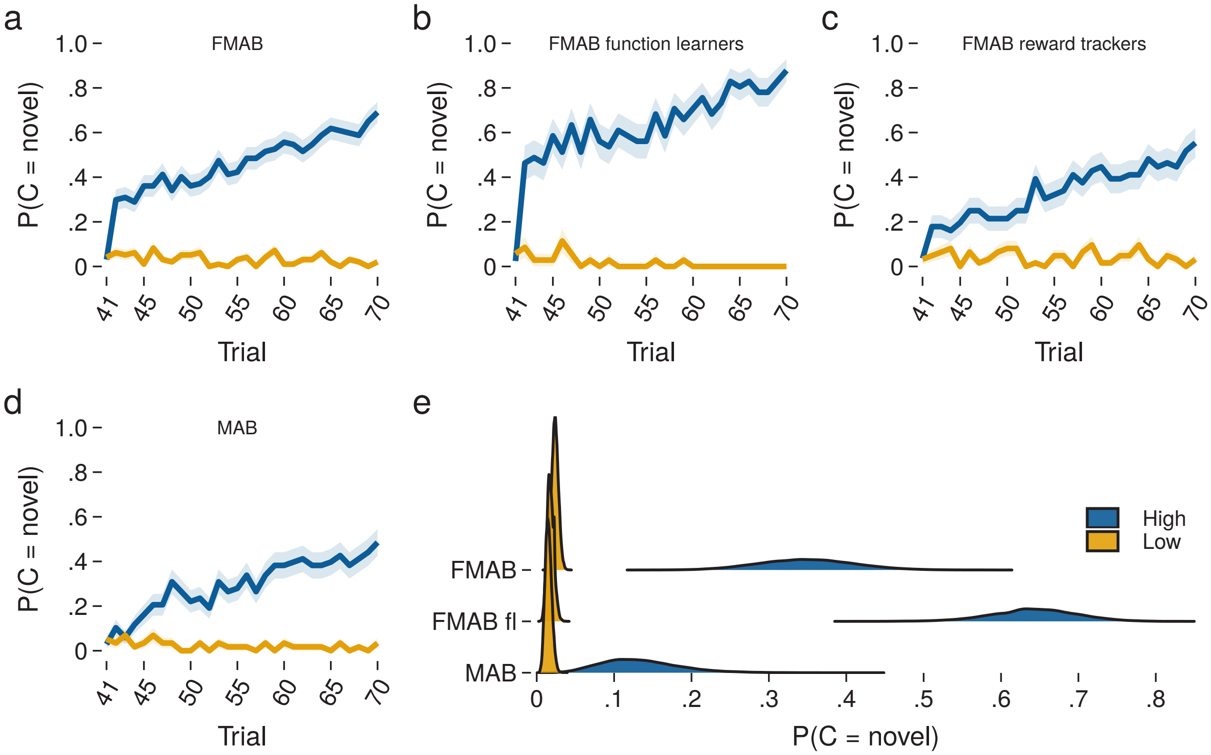


Figure 5. Proportions of choices allocated to the novel option from trial 41 onwards in Experiment 1. (**a**) From trial 42 onwards, participants in the FMAB high value condition choose the novel option more often than those in the FMAB low value condition. (**b**) The same pattern, but with stronger magnitude, is evident when only looking at function learners in the FMAB conditions. (**c**) Reward trackers in the FMAB conditions choose both the high and low value option in similar proportions at the beginning, but as predicted by the model simulations, they learn over time and tend to choose the high options more frequently and the low value options less frequently later on. (**d**) Participants in the MAB conditions make choices similar to reward trackers in the FMAB conditions. In all four figures, lines reflect average choice proportions across participants, while filled bands indicate the standard errors of the means. (**e**) Densities of posterior distributions over the probability of choosing the high or low value novel option in trials 41 to 70, estimated with a Bayesian binomial model. Distributions show a clear difference for both FMAB conditions as a whole and for the function learning subgroup. In contrast, the difference between the MAB conditions is substantially smaller.

model predictions, FMAB participants – and function learners in particular – should estimate the high value novel option to have higher reward than the low value option on trial 41 (Figure 4a, top left panel). By contrast, following the BH-UCB model's predictions, MAB participants and reward trackers in the FMAB condition should evaluate both novel options as roughly equal (Figure 4a, top right panel). As expected, on trial 41 FMAB participants correctly estimated the value of the high value option (19.05) to be higher than the low value (16.06) option ($BF^*_{10} = 30.86$), while there was moderate evidence that the estimates of the MAB participants were equal (17.71 points in the high and 17.78 in the low condition, $BF^*_{01} = 5.22$, non-directional H_1). Function learners exhibited an even stronger effect, estimating the value of the high value option (21.88) to be higher than the low value option (14.37, $BF_{10} = 4918$), with estimates being closer to the true values of 6.5 and 33.5 points (Figure 6a, trial 41 panel). By contrast, reward trackers behaved similarly to MAB participants, producing moderate evidence that their estimates for both options were equal (16.99 in the high value and 17.02 in the low value condition, $BF_{01} = 5.10$, non-directional H_1). By trial 70, as predicted by the BH-UCB model simulations, MAB participants and reward trackers learned about the novel options and exhibited the predicted differences in valuations of the high and low value novel options (MAB: 24.79 and 15.72, $BF_{10} = 5.74 \times 10^8$; FMAB reward trackers: 24.91 and 15.13, $BF_{10} = 9.1 \times 10^7$; Figure 6a, trial 70 panel). The difference in valuation between the two options increased further for the

($BF^*_{10} = 2.3 \times 10^9$ and $BF^*_{10} = 3.6 \times 10^{60}$; intercept-only model as H_0) and the MAB condition ($BF^*_{10} = 4.5 \times 10^{11}$ and $BF^*_{10} = 3.0 \times 10^{37}$; intercept-only model as H_0).

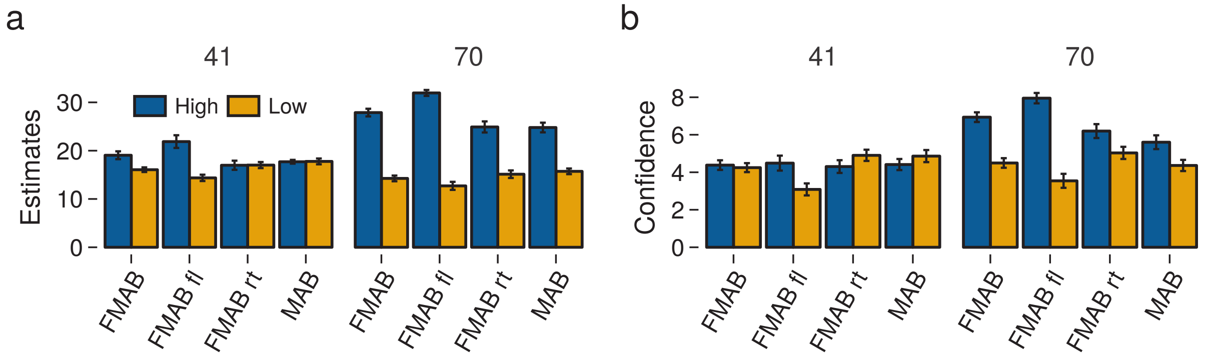


Figure 6. Estimated reward of the novel option and associated confidence in Experiment 1 from the estimation task on trial 41 and 70. (**a**) Estimated reward of the novel option. On trial 41, participants in the FMAB conditions, and function learners (FMAB fl) in particular, expressed correct beliefs that the high valuenovel option is more rewarding than the low option. MAB participants and reward trackers (FMAB rt) estimated both novel options to have an equal value. (**b**) Confidence in the estimated reward of the novel option. On trial 41, function learners (FMAB fl), but not all FMAB participants, were more confident about the high value novel option than the low option. Error bars in (**a**) and (**b**) are standard errors of the means, while 41 and 70 are panels showing the estimation task's data in those trials.

FMAB participants in general and function learners in particular (FMAB: 27.89 and 14.26, $BF_{10} = 2.2 \times 10^{27}$; FMAB function learners: 31.95 and 12.71, $BF_{10} = 8.6 \times 10^{26}$).

The GP-UCB model predicts that function learners' confidence in their reward estimation on trial 41 will be higher for the high value novel option than the low value option (Figure 4a, middle left panel). This expectation did not hold for all FMAB participants ($BF^{*}_{10} = 0.21$), but was confirmed for participants classified as function learners, who indicated higher confidence for the high value option (4.49) than the low value option (3.09), $BF_{10} = 9.68$ (Figure 6b, trial 41 panel). By contrast, following the predictions of the BH-UCB model, confidence was expected to be equal for both novel options for MAB participants and reward trackers (Figure 4a, middle right panel). Indeed, there was moderate evidence that participants in the MAB conditions had equal confidence ratings (high vs. low: 4.41 vs. 4.86), $BF^{*}_{01} = 3.25$ (non-directional H_1). Reward trackers in the FMAB conditions showed a pattern resembling the MAB participants, providing confidence ratings for the two options that were close to each other (high vs. low: 4.30 vs. 4.90), with weak to moderate evidence that they were equal ($BF_{01} = 2.31$, non-directional H_1). On trial 70, as predicted by simulations of both models, confidence in the high value option increased and was higher than for the low value option (Figure 6b, trial 70 panel), more so for the FMAB participants (6.94 and 4.49, $BF_{10} = 7.5 \times 10^{7}$) and function learners (7.95 and 3.54, $BF_{10} = 8.6 \times 10^{11}$) than MAB participants (6.20 and 5.03, $BF_{10} = 2.39$) and reward trackers (5.60 and 4.36, $BF_{10} = 3.60$).

Discussion

Experiment 1 produced evidence for the functional generalization effect. Participants in the FMAB condition, and those relying on a function learning strategy in particular, avoided the novel option in the low value condition and chose it more frequently in the high value condition. Participants' beliefs about expected rewards further corroborated this result – they correctly believed that the reward of the high value option was higher than that of the low value option. Moreover, participants were more confident in their predicted rewards in the high value than in the low value condition, as again predicted by our model. By contrast, participants in the MAB conditions and FMAB participants adopting a reward tracking strategy were not able to distinguish between the novel options. Their beliefs about the expected rewards and their confidence in those beliefs did not differ between the two options. Consequently, while they eventually discovered the high value option, they did so much later than FMAB participants and particularly function learners.

A key preregistered hypothesis concerned the 41st trial, where we expected FMAB participants to choose the high value option more frequently than the low value option. This prediction was not confirmed by the results, but we found a strong effect on the 42nd trial, as well as when comparing all trials from the 41st trial onwards. Moreover, data derived from the estimation task underpinned these results further. Why did participants only start considering the novel option after the 41st trial? By the end of trial 40, many participants might have already settled on their next choice, registering it shortly after the feedback. At that point, the novel option would still have been half-transparent, likely hindering its detection. After the estimation task asked participants to reflect on the value of all options, both old and novel, the expected effects did occur. The extent to which this can be attributed to reflection elucidated by the estimation task, or to noticing the novel option for the first time during the estimation task, is unclear. We will turn to identifying whether attention or reflection underpins the observed effects from 42nd trial onwards in Experiment 3.

To conclude, we found clear evidence for the predicted functional generalization effect. Our results show the generative potential of theories of function learning, and suggests a previously unrecognized mechanism explaining diverging reactions towards novel options. Beyond that, our theory takes a probabilistic approach to function learning, enabling us to predict people's confidence in their expectations regarding how rewarding options are. Participants' confidence in their predictions has received little attention in previous category and function learning studies, but in a decision context, where an agent can choose what to observe and learn about, confidence can be invaluable (Boldt, Blundell, & De Martino, 2019; Folke, Jacobsen, Fleming, & De Martino, 2017). We found evidence that function learners maintain a measure of the uncertainty in their knowledge. When the goal is to maximize rewards, choices are biased towards highly rewarding options, so that relatively more information is obtained for more rewarding options, resulting in more uncertainty in lower rewarding regions. These findings pave the way for Experiment 2, where we examined whether people eagerly approach more uncertain options.

Experiment 2: Functional uncertainty guidance

The second preregistered experiment investigated functional uncertainty guidance, i.e. whether people explore options to improve their functional knowledge, thereby preferring novel options with higher predictive uncertainty. We again used a between-subjects design. One group was assigned a novel option with feature values from within the experienced range. We will refer to this group as the *FMAB ordinary-novel* condition. Another group was assigned a novel option with feature values from outside the experienced range. We will refer to this group as the *FMAB exotic-novel* condition. We selected features such that the novel options in both conditions had exactly the same reward in expectation. However, uncertainty about the value of the exotic-novel option was expected to be perceived as higher, since it had feature values from outside the experienced range and so was less similar to the old options than the ordinary-novel option.

As in Experiment 1, we used an MAB version of the task as a control condition. Since both types of novel option had the same expected reward, there was no differentiation between the novel options in the MAB version. Thus, a single *MAB* condition sufficed.

Method

We recruited 423 participants (207 female, $M_{\text{age}} = 37.4$ and $SD_{\text{age}} = 10.9$) through Amazon's Mechanical Turk using the same eligibility requirements as in Experiment 1. There were 182 participants in the FMAB exotic-novel condition, 180 participants in the ordinary-novel condition, and 61 participant in the MAB condition. We followed the same sampling plan as in Experiment 1 and stopped after we had reached our budgetary limit (Appendix A). We rewarded participants with a fixed payment of $0.70 and a performance-dependent bonus of $1.40 on average. The experiment took 12.6 minutes on average. The study was approved by the UCL Research Ethics Committee.

The task in Experiment 2 was the same as in Experiment 1, with the only difference being how the novel options were constructed. In the FMAB *ordinary-novel* condition we set the novel option's features to $x_{1,10} = 0.33$ and $x_{2,10} = 0.34$, making it similar to the already experienced options (i.e. the feature values were within the $\mathcal{U}(.25, .35)$ range from which feature values of old options were drawn), yielding a medium expected reward of 25 points. In the FMAB *exotic-novel* condition, the novel option had feature values from outside the experienced range, $x_{1,10} = 0.01$ and $x_{2,10} = 0.98$. Crucially, the expected reward of this option was again 25 points.

The procedure was exactly the same as in Experiment 1. Participants who had participated in Experiment 1 were not allowed to participate in Experiment 2.

Predictions

We generated predictions by simulating the GP-UCB and BH-UCB model and preregistered them before data collection commenced (Stojic et al., 2018b). The results of the simulation show that after introduction of the novel option on the 41st trial, the GP-UCB model chooses the exotic-novel option with a higher probability than the ordinary-novel option (Figure 4D, left panels). The GP-UCB model learns the underlying reward function during the first 40 trials, allowing it to correctly predict the mean reward of the ordinary-novel option, while it underestimates the mean reward of the exotic-novel option. However, as predictions for the exotic-novel option are more uncertain than predictions for the ordinary-novel option, more information can be gained from choosing the exotic novel option. The UCB rule takes into account the informativeness of choices, and here the difference in uncertainty outweighs the difference in predicted reward, resulting in a small but reliable preference for the exotic-novel option compared to the ordinary-novel option. As the uncertainty about reward reduces with experience, this relative preference disappears within 5 trials.

The predicted relative preference for the exotic-novel option rests on both ingredients of the GP-UCB model: function learning and uncertainty-guided exploration. A sophisticated reward-tracking strategy which also takes into account rewards and uncertainty, such as instantiated by the BH-UCB model, is not enough. The simulation of the BH-UCB model shows that it allocates the same proportion of choices to both types of novel option (Figure 4d, right panels). By ignoring the feature values, this model is unable to differentiate between the novel options a priori, assigning the same uncertainty to both. This shows that the underlying representation from which uncertainty is derived matters. If options are represented as in the BH-UCB model, being drawn from a common distribution but otherwise independent, this results in different uncertainty than if options' reward is represented as a function over feature values, consequently leading to diverging choices. The functional uncertainty effect, where an exotic-novel option is chosen more often than an ordinary novel option because predictions are more uncertain for the former, is a new prediction directly derived from our function learning account of experiential decision making.

The predicted difference in choice proportions between the exotic-novel and ordinary-novel option is not as large as the predicted difference between the high value and low value novel options in Experiment 1 (Figure 4). Looking at the expected reward of the novel options, we can see that this is mostly due to the GP-UCB model underestimating the reward for the exotic-novel option as compared to the ordinary-novel option. This is a direct consequence of using an RBF kernel – predictions outside the experienced feature space tend to reverse back to the overall prior mean value. As indicated earlier, we do not make a strong theoretical commitment to the RBF kernel. Participants may not solely rely on a mechanism of generalization which mirrors this kernel. Instead, they may extrapolate by either assuming longer-distance dependencies (A. G. Wilson et al., 2015), employing rule-based learning (Busemeyer et al., 1997) or compositional learning (Schulz et al., 2017). Such alternative learning mechanisms would yield a smaller difference in predicted rewards whilst leaving the uncertainty difference relatively intact, leading to an increased relative preference for the exotic-novel option. As such, the difference between the exotic-novel and the ordinary-novel option could turn out to be larger than in our preregistered simulation.

As in Experiment 1, the GP-UCB predictions should hold for function learners in the FMAB conditions, and the FMAB conditions as a whole if enough participants are indeed function learners. The predictions of the BH-UCB model should hold for the MAB conditions and reward trackers in the FMAB conditions.

Results

Choice proportions. Having preregistered the hypotheses for Experiment 1 and 2 simultaneously, the primary hypothesis for the second experiment again concerned the 41st trial. As in Experiment 1, there was again no evidence for the expected difference in choice proportions on the 41st trial, neither for participants in the FMAB conditions (6.6% in exotic-novel and 6.7% in ordinary-novel condition, BF^*_{10} = 0.07), nor for function learners (14.3% in exotic-novel and 8.1% in ordinary-novel condition, BF^*_{10} = 0.25). In the MAB condition, 4.9% of the participants chose the novel option on trial 41.

Since the procedure was the same as in Experiment 1, the same issues with participants' attention to the novel option arose in Experiment 2. We therefore proceeded to explore participants' behavior from the 42nd trial onwards. On the 42nd trial, participants in the FMAB exotic-novel condition chose the novel option more frequently (18.1%) than participants in the ordinary-novel condition (7.2%), BF_{10} = 11.74. According to the GP-UCB model simulations, this difference was predicted to vanish rapidly. The behavioral data shows differences in choice proportions ranging from 4% to 12% over a longer period until trial 52, when the differences disappear. This relative preference over a prolonged period of time is likely due to participants exhibiting slower learning rates than the GP-UCB model (Figure 7a). However, for most of these trials, statistical evidence of a difference was relatively weak, with $BF_{10} < 3$.

Based on performance in the functional knowledge task, we classified 144 out of 362 participants (39.8%) in the FMAB conditions as function learners, a similar proportion as in Experiment 1. On the 42nd trial, function learners showed a similar preference for the exotic-novel option (25.7%) compared to the ordinary-novel option (12.2%) as FMAB participants more generally, however resulting in a smaller Bayes factor of BF_{10} = 1.36, likely due to the smaller sample size. As predicted by our models, on trial 70 there was no difference between the conditions in how often participants chose the novel option, ranging from 12.2% in the FMAB ordinary-novel condition to 19.7% in the MAB condition (BF_{01} = 31.7).

As a final analysis of the choice data, we again estimated the probability of choosing the novel option across more than a single trial using a Bayesian binomial model. As model simulations predicted effects to diminish over trials, we focused our analysis on the first 15 trials after introduction of the novel option (i.e. trial 41 to 55). The posterior distributions over the probability of choosing the novel option showed a small but consistent separation between the FMAB conditions (Figure 7e). For the exotic-novel condition the median was 9.3% (95% CI [6.5, 12.6]) and for the ordinary-novel condition it was 6.9% (95% CI [5.1, 9.0]), with 91% of the samples from the posterior distribution for exotic-novel condition being larger than the ordinary-novel estimate and a mean difference of 2.5%. Repeating the same analysis for function learners only showed a smaller difference – the median for the exotic-novel condition was 12.7% (95% CI [7.4, 19.4]) and 10.6% (95% CI [6.9, 14.9]) for the ordinary-

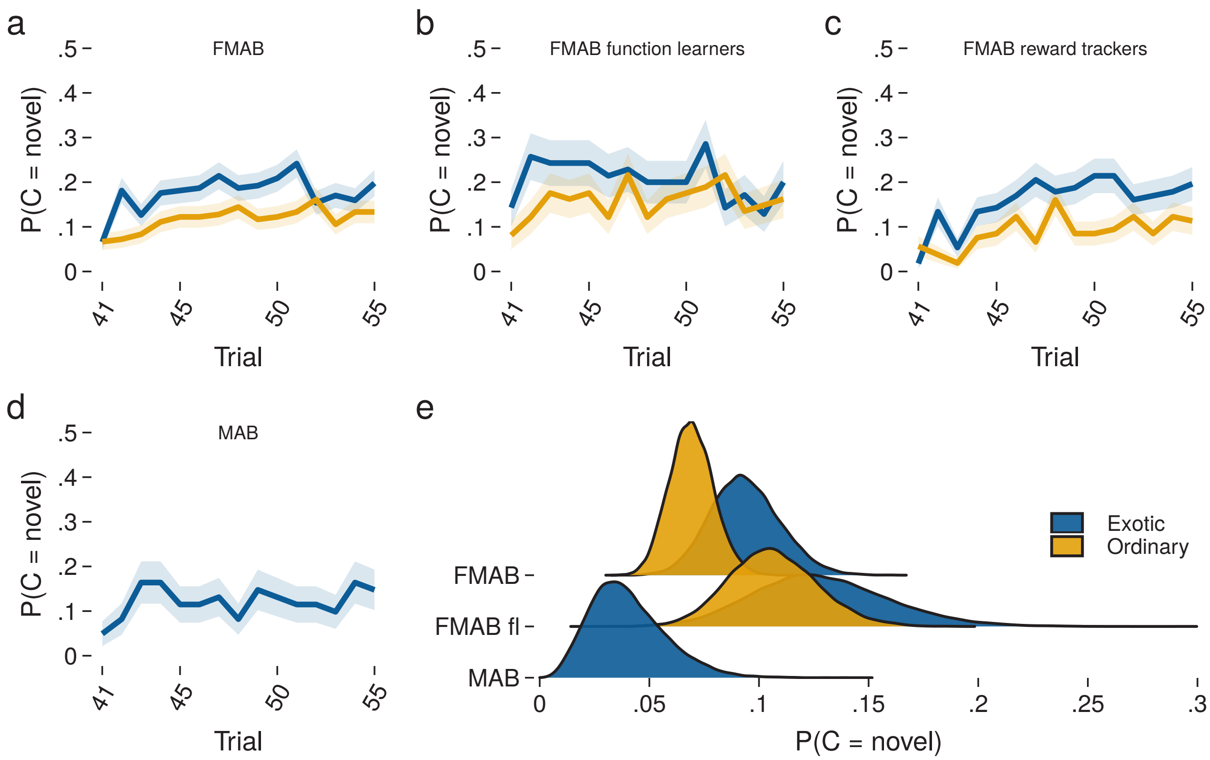


Figure 7. Proportions of choices allocated to the novel option from trial 41 to trial 55 in Experiment 2. (**a**) Participants in the FMAB exotic-novel condition start choosing the novel option more often than participants in the FMAB ordinary-novel condition on trial 42, but as predicted by the model simulations the difference starts decreasing soon after that, from trial 52 onwards. (**b**) Function learners show a similar pattern, with a greater difference in choice proportions. Note that choice proportions are noisier due to a smaller number of participants. (**c**) By contrast, reward trackers in the FMAB conditions initially choose both novel options in similar proportions, starting with low allocations to novel options and increasing slowly. (**d**) Participants in the MAB conditions make choices similar to reward trackers in FMAB conditions. In all four figures lines are mean proportions across participants, while bands are standard errors of the means. (**e**) Densities of posterior distributions of probabilities of choosing the novel option in trials 41 to 55 in the FMAB exotic-novel and ordinary-novel conditions and MAB novel condition, estimated by a Bayesian binomial model. The estimated posterior probabilities show a small but robust difference for both FMAB conditions as a whole and the function learning subgroup, while the estimated probability for the MAB condition is substantially smaller.

novel, with 73% of samples for the exotic-novel condition larger than the ordinary-novel estimate and a mean difference of 2.2%. By contrast, for participants in the MAB condition the median probability of choosing the novel option was 3.8% (95% CI $[1.3, 7.9]$), with 98% of the samples for the FMAB exotic-novel condition being larger than those for the MAB condition and a mean difference of 5.4%.

Expected rewards and uncertainty. Next, we examined whether participants' beliefs elicited in the estimation task aligned with their choices (Figure 8a and 8b).[4] Participants in the FMAB conditions estimated the ordinary-novel option to have a higher mean reward (18.54) than the exotic-novel option (15.75), $BF_{10} = 24741$. This was in line with the GP-UCB model's predictions. This difference was even larger when focusing solely on function learners (19.49 for the ordinary-novel and 15.37 for the exotic-novel option, $BF_{10} = 1637$). For reward trackers, there was no evidence that estimates of their expected rewards for the exotic-novel option (17.89) were equal to those for the ordinary-novel option (15.99), $BF_{01} = 0.20$, non-directional H_1.

As expected, participants in the FMAB conditions were less confident in their predictions for the exotic-novel option (3.84) than for the ordinary-novel option (4.74), $BF^*_{10} = 80.54$. This difference was again larger for function learners (3.13 for the exotic-novel and 4.96 for the ordinary-novel option, $BF_{10} = 2291$). For reward trackers, there was moder-

[4] The estimated mean values and confidence ratings again passed our preregistered sanity checks, both in the FMAB conditions ($BF^*_{10} = 6.6 \times 10^{25}$ and $BF^*_{10} = 6.6 \times 10^{101}$) and in the MAB conditions ($BF^*_{10} = 46$ and $BF^*_{10} = 6.4 \times 10^{17}$).

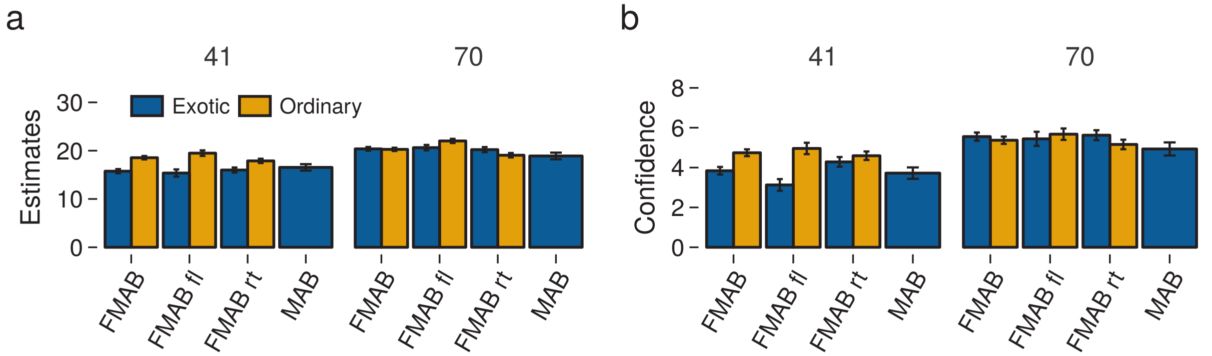


Figure 8. Estimated reward of the novel option and associated confidence in Experiment 2 for the estimation task on trial 41 and 70. (**a**) Estimated average reward of the novel option. FMAB participants and function learners in FMAB conditions (FMAB-fl) underestimate the exotic-novel option on trial 41 but the difference disappears by trial 70. Reward trackers (FMAB-rt) also underestimate the exotic-novel option on trial 41, but not on trial 70. (**b**) Confidence in the estimated reward of the novel option. As predicted, FMAB participants and function learners are less confident about the exotic-novel option than the ordinary-novel option on trial 41. This difference disappears by trial 70. Error bars in (**a**) and (**b**) are the standard errors of the means, while 41 and 70 are panels showing the data from the estimation task on those trials.

ate evidence that their confidence ratings for the exotic-novel option (4.59) were equal to the ratings for the ordinary-novel option (4.29), BF_{01} = 4.43, as predicted by our model simulations.

Participants' confidence for the exotic-novel option increased substantially from trial 41 (3.84) to 70 (5.55), BF_{10} = 7.3×10^6, indicating that once they had tried out the exotic-novel option, their knowledge of its value improved. As predicted by our model simulations, the differences between the FMAB conditions disappeared by the end of the task, in both estimated expected rewards (20.37 in the exotic-novel and 20.27 in the ordinary-novel condition, BF_{01} = 8.45, non-directional H_1) and confidence ratings (5.55 in the exotic-novel and 5.37 in the ordinary-novel condition, BF_{01} = 6.97, non-directional H_1; trial 70 in Figure 8a and 8b).

It is conceivable that the observed patterns in estimated rewards, confidence, and choices might not hold on the level of individual participants, but were due to averaging in the conditions. This could occur if a subset of participants in the exotic-novel condition shows the expected pattern in the estimation task (lower estimated value and lower confidence) but actually *does not* choose the novel option, while another subset does not show this pattern (for instance valuing the novel option higher than any of the old options), but does choose the novel option. We performed an additional analysis to determine whether this was the case. For each participant in the FMAB exotic-novel condition, we calculated the difference in estimated value and confidence between the subjectively best old option (the old option which received the highest rating in terms of expected reward) and the novel option. For participants who believed that the novel option had a higher value than any of the old options, this difference would be negative for their estimated value. Similarly, for participants who were more confident in their estimation of the novel option than any of the old options, this difference would be negative for their rated confidence. Looking at those participants who chose and did not choose the novel option on trial 42, we found that both groups showed positive difference scores. Participants who chose the novel option rated its value as lower than an old option (mean difference 9.27, SE = 1.59; BF_{10} = 2.03×10^4, H_0 smaller than or equal to zero) and also had lower confidence in their estimation for the novel option than the best old option (mean difference 3.70, SE = 0.65; BF_{10} = 1.5×10^4). Participants who did not choose the novel option on trial 42 showed the same pattern, reporting that the novel option was of lower value than the best old option (mean difference 9.56, SE = 0.47; BF_{10} = 1.21×10^{42}) and that they had lower confidence in their rating of the novel option than the best old option (mean difference 3.40, SE = 0.25; BF_{10} = 7.1×10^{24}). This confirms that the uncertainty-guidance effect is not an artifact of aggregating individual data.

Discussion

Experiment 2 revealed moderate evidence that participants preferred the exotic-novel over the ordinary-novel option in the period soon after the novel option was introduced. Analyzing choice proportions, we found a moderate preference for the exotic-novel option on the 42nd trial, and a small overall preference in the period from the 41st to the 55th trial. Similar to Experiment 1, our hypothesis concerning the 41st trial was not confirmed. Importantly, participants' beliefs about average rewards and their confidence in these beliefs provided further evidence for functional uncertainty guidance: exploration of novel options to gain functional knowledge.

Beyond that, our findings more closely followed the simulations with a Radial Basis Function kernel than what would have been expected if participants had extrapolated more linearly (e.g. Busemeyer et al., 1997; Hoffmann, von Helversen, & Rieskamp, 2016). Using an RBF kernel, our model predicted a relatively small difference in choice proportions between the exotic-novel and ordinary-novel options, but the difference could have been larger if extrapolation relied on a different kernel which does not underestimate the average reward of the exotic-novel option. While our previous research showed that a similarity-based kernel such as the RBF kernel describes participants' learning well (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojic, 2016), that evidence was not sufficiently strong to discard the possibility that humans also incorporate linear extrapolation. The finding that participants did indeed underestimate the value of the exotic-novel option suggests that their function learning was predominantly driven by a similarity-based representation of the function.

While our results indicate that participants integrated uncertainty into their decision process, it may have played a less prominent role than in our models, leading to moderate rather than strong differences in choice proportions between the exotic and the ordinary novel option. Indeed, contrary to predictions from the BH-UCB model, participants rarely ever chose the novel option in the MAB condition. A decreased probability of choosing the novel option can be accounted for by relatively smaller values of the β parameter in the UCB choice rule, which would dampen the exploration of uncertain options. However, it is unclear whether this also holds for the FMAB conditions. Since past research has shown that participants can adapt the extent to which they rely on uncertainty to the encountered environments (e.g. Behrens, Woolrich, Walton, & Rushworth, 2007), it is likely that uncertainty guidance played a more important role in the FMAB condition where directed exploration can lead to better knowledge about the underlying function and thereby improve future decisions.

Experiment 3: Attention and reflection

Given the results of Experiment 1 and 2, the following question remained: why did the predicted effects only occur after the 41st trial, and not – as expected – on the 41st trial? At least two interpretations are conceivable. First, this could have been an attention effect: participants may have already settled on choosing an old option before the start of the trial and therefore did not notice the slowly appearing novel option on trial 41. Second, the estimation task could have triggered further reflection on all options, including the novel one, such that participants only decided to choose or ignore the novel option after explicitly evaluating it.

One source of evidence for what might have driven this effect is to look at reaction times. We therefore analyzed the time participants took to make a choice in trials around the 41st trial in Experiment 1 and 2 to glean initial evidence about these two hypotheses. In the five trials preceding the 41st trial, FMAB participants were making choices typically in under a second (Experiment 1: 0.73s, SE = 0.04; Experiment 2: 0.79s, SE = 0.03). Participants' choice times were similar on the 41st trial (Experiment 1: 0.85s, SE = 0.07, BF_{01} = 2.3; Experiment 2: 0.86s, SE = 0.05, BF_{01} = 5.3, non-directional H_1; excluding the estimation task time), which was indeed too fast for the novel option to become fully visible (since this took 3s in total). By contrast, participants took much longer to choose an option on the 42nd trial relative to the five trials preceding the 41st trial (Experiment 1: 1.87s, SE = 0.12, $BF_{10} = 3.8 \times 10^{26}$, non-directional H_1; Experiment 2: 1.84s, SE = 0.08, $BF_{10} = 1.9 \times 10^{33}$, non-directional H_1). In comparison to the 42nd trial, participants speeded up again in the five subsequent trials (Experiment 1: 0.79s, SE = 0.04, $BF_{10} = 1.3 \times 10^{32}$, non-directional H_1; Experiment 2: 0.81s, SE = 0.03, $BF_{10} = 4.9 \times 10^{34}$, non-directional H_1). The increase in choice time on the 42nd trial could be a result of considering to choose the novel option, but it could also be due to the estimation task that appeared on the previous trial. For brevity we do not report results for the MAB conditions in either experiment, which were qualitatively similar, and we collapsed the analysis across FMAB conditions, as results were also similar.

These choice time analyses seem to suggest that the one-trial delay of the expected effects in Experiments 1 and 2 was due to participants' quick decision making and failing to attend to the novel option. However, these results are inconclusive on their own. Participants could have become aware of the novel option because it was fully visible during the estimation task. But the estimation task itself could also have caused them to further reflect on all options, which in turn led them to either choose or avoid the novel option. We tried to further disentangle these explanations in Experiment 3. In Experiment 3a and 3b, we examined whether drawing participants' attention more clearly to the appearance of the novel option is sufficient to remove the delayed response to novelty, which would provide support to the attention explanation. In Experiment 3c, we examined whether further reflection on all options is also necessary, which would provide support for the reflection hypothesis.

We focused on the FMAB high and FMAB low value conditions from Experiment 1. The expected and observed effects were strongest for these conditions, allowing us to efficiently test our hypotheses. As before, we preregistered all of the experiments on the OSF website: Experiment 3a at `https://osf.io/h5uqr/` (Stojic, Schulz, Analytis, & Speekenbrink, 2019a), Experiment 3b at `https://osf.io/37ayn/` (Stojic, Schulz, Analytis, & Speekenbrink, 2019b) and Experiment 3c at `https://osf.io/tg5kc/` (Stojic, Schulz, Analytis, & Speekenbrink, 2019c).

Method. We recruited 419 participants (170 female, $M_{\text{age}} = 37.8$ and $SD_{\text{age}} = 11.8$) through Amazon's Mechanical Turk using the same eligibility requirements as in Experiment 1 and 2. There were 117 participants in Experiment 3a (57 in the FMAB high value and 60 in the low value condition), 181 participants in Experiment 3b (90 in the FMAB high value and 91 participants in the low value condition)[5] and 121 participant in Experiment 3c (61 in the FMAB high value and 60 participants in the low value condition).

We followed the same sampling plan as in Experiment 1 and 2 (Appendix A). We rewarded participants with a fixed payment of $1.00 and a performance-dependent bonus of $1.60 on average. The experiments took 13.2 minutes on average. All studies were approved by the UCL Research Ethics Committee.

Methods and procedure for Experiment 3 were similar to the two FMAB conditions in Experiment 1, except for some key modifications. One modification concerned all three experiments. We implemented a period of two seconds in which participants could not choose an option at the start of each trial and this period was clearly marked by surrounding all options with a thick black border. Participants could register their choice only after this period had ended, which was marked by the removal of the borders. This was done to prevent rapid choices and increase the chance of noticing the novel option appearing. Other modifications were specific to each experiment, and concerned the way in which the novel option was introduced, and the presence or absence of the estimation task.

In Experiment 3a, instead of slowly becoming opaque, we made the novel option appear on trial 41 by flickering four times over a period of 1 second. In addition, the text below the options changed into a simple message stating "A new option has been added. Everything else about the task is the same as before and all options will remain available until the end of the game." These modifications were designed to make the introduction of the novel option highly noticeable. Moreover, there was no estimation task in trial 41 and 70, so participants were not explicitly invited to reflect on all options. This experiment was thus designed to test whether increased attention to the novel option alone would be sufficient to induce approaching or avoiding the novel option.

The results of Experiment 3a suggested that strongly directing participants' attention to the novel option might have had unanticipated effects on functional generalization. Hence, in Experiment 3b the novel option appeared on trial 41 by slowly becoming opaque over a period of time, similar to Experiment 1 and 2. We reduced the time of this fading-in process from 3 seconds to 2 seconds, to match it to the period in which participants could not register a response. As in Experiment 3a, there was no estimation task in trial 41 and 70. This experiment was designed to assess how a more subtle way to draw participants' attention to the novel option affects their feature-based choices.

Finally, in Experiment 3c, we assessed how making participants explicitly reflect on each option influenced their feature-based choices of the novel option. In this experiment, the estimation task appeared at the beginning of trial 41 before participants made a choice, rather than after the choice as in Experiment 1 and 2. Crucially, the novel option was simply added to the estimation task, without any accompanying visual effects.

In addition to these larger modifications, we also made several minor changes. We clarified the instructions further, mainly in the attempt to improve readability and style, and increased the payoffs slightly. In Experiment 3b and 3c, we also added six questions at the end of the experiment, alongside demographic questions, to probe participants' knowledge of the reward function and the appearance of the novel option. In each of the three experiments, we followed the same sampling plan as in Experiment 1, but with a budget stopping rule of $300. Finally, participants who had participated in Experiment 1 and 2 were not allowed to participate in Experiment 3.

All model-based predictions for Experiment 3a, 3b, and 3c were identical to those for Experiment 1 (Figure 4a).

Results

In Experiment 3a, the novel option appeared in a salient flickering manner in an attempt to ensure that participants noticed it. A high proportion of participants chose the novel option on the 41st trial in both the high value (63.2%) and the low value (70%) condition. These proportions were much higher than what we had found in Experiment 1 on the 42nd trial (30% and 6%). There was no evidence for the expected difference in choice proportions allocated to the novel option on the 41st trial (63.2% in high value and 70% in low value condition, $BF^*_{10} = 0.29$). Same holds for 64% of participants that were classified as function learners (76.5% chose novel option in high value and 73.2% in low value condition, $BF^*_{10} = 0.26$). Same as in Experiment 1, here we found the predicted difference after the 41st trial (Figure 9a); however, this was largely due to participants in the low value condition substantially decreasing the frequency of choosing the novel option.

The results of Experiment 3a revealed that strongly directing participants' attention to the novel option can affect the way in which they use the option's features to guide their choices. We found no evidence for the expected difference between the high and low value condition on the 41st trial. Instead, many more participants chose the novel option in Experiment 3a than in Experiment 1, independent of its fea-

[5] We added 54 participants from a pilot, details can be found in the preregistration document (Stojic et al., 2019b).

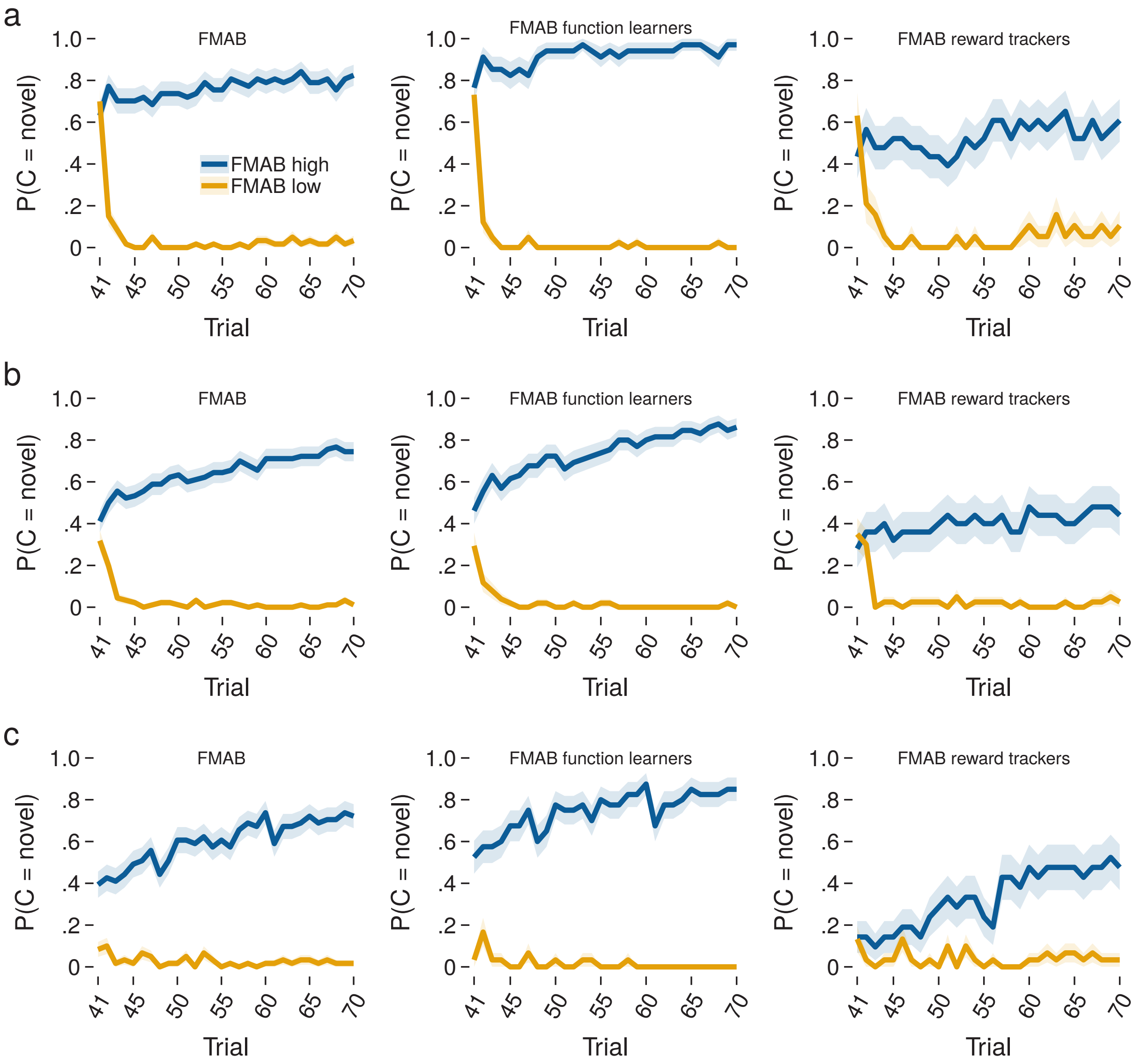


Figure 9. Proportions of choices allocated to the novel option from trial 41 onwards in Experiment 3a (**a**), Experiment 3b (**b**) and Experiment 3c (**c**). Left column are the results for the FMAB condition, while middle and right column are results of the FMAB conditions decomposed into function learners and reward trackers. In all figures, lines are average choice proportions across participants, while filled bands are standard errors of the means.

tures. It is possible that participants interpreted the flickering as indicating that the novel option was in some sense "special" or different to the old options beyond the differing feature values. They could have also interpreted the flickering as an additional feature. In either case, this means the feature-reward function governing the old options would not apply to the novel option. This would increase the uncertainty about the novel option's reward, likely making it an attractive choice regardless of the value of the two original features. Since such effects are outside the scope of our current theory, in Experiment 3b we considered a subtler approach to ensuring participants noticed the novel option upon its introduction. We reasoned that the original way of introducing the novel option in Experiment 1 and 2 would reduce the chances of interpreting the visual effects accompanying the appearance of the novel option as a new feature. However, by reducing the time of the fade-in to two seconds, equaling the time of the forced holdout in which participants could not register their choice, the chance of failing to notice the novel option should also be low.

In Experiment 3b, the proportions of choosing the novel option indeed decreased in comparison to Experiment 3a:

on the 41st trial 41.1% of participants chose the novel option in the high value and 31.9% in the low value condition. While in the expected direction, there was only negligible evidence for the expected difference between the conditions (BF^*_{10} = 0.4). Function learners (64% of participants) showed a stronger difference (46.2% in the high value and 29.4% in the low value condition), but there was only weak evidence for the expected effect (BF^*_{10} = 1.18). Similar to Experiment 3a, the difference was more substantial after the 41st trial (Figure 9b). These results indicate that bringing the novel option to attention in a subtler way was not sufficient to generate the expected effect on the 41st trial. We therefore tested if further evaluation of all options was necessary for the expected effect in our final Experiment 3c.

Experiment 3c examined whether introducing the estimation task before participants made their choice on the 41st trial would bring about the expected effect. Here, we found convincing evidence for the expected difference between the conditions. On the 41st trial, 39.3% in the high value condition and 8.3% in the low value condition chose the novel option, $BF^*_{10} = 731$. As expected, function learners (58% of participants) showed a larger difference: 52.5% in the high value and 3.3% in the low value condition chose the novel option ($BF^*_{10} = 1.2 \times 10^4$). This pattern was followed by a further increase in difference in the remaining trials (Figure 9c). [6]

Discussion

Experiment 3 provided further clarification on why the effects predicted by our theory did not emerge on the 41st trial in Experiment 1 and 2. Specifically, we put forward and assessed two possible explanations. The first was that participants failed to notice the appearance of the novel option before registering their choice. The second was that the additional reflection and evaluation induced by the estimation task made them attend to the novel option and realize its value, thereby affecting their choice. We found support for the second explanation. The estimation task made participants not only aware of the novel option, but also required them to explicitly reflect upon the options' values. This evidence suggests that our theory is supported when participants are aware of the novel option and reflect on the options in the consideration set before making a choice.

The manipulations to increase participants attention to the novel option of Experiment 3a and 3b were not sufficient to produce the expected effect. While the more subtle introduction in Experiment 3b produced an effect in the expected direction, no statistically meaningful differences emerged. The salient flickering visual effect in Experiment 3a on the other hand led the large majority in both conditions to choose the novel option, irrespective of whether the feature values indicated high or low rewards. It seems that the visual effect not only drew attention to the novel option, but may have also been interpreted as a new feature or a change to the feature-reward function altogether, thereby increasing the uncertainty of the option's reward which overwhelmed the functional generalization effect. It is possible that visual effects in Experiment 3b, despite being subtler, were still interpreted in that manner, albeit less strongly. Such effects of salience and attention are currently beyond the scope of our GP-UCB model, which assumes perfect attention to all options and that a novel option is perceived as "just another option" governed by the same feature-reward function. Alternatively, attention per se can bias choice (Krajbich, Armel, & Rangel, 2010; Shimojo, Simion, Shimojo, & Scheier, 2003) and recent work showed that attending to an option can even amplify its value (Smith & Krajbich, 2019). Hence, exogenously drawing attention to the novel option and thereby diminishing attention for all other options could explain results in Experiment 3a and 3b, without any effect at the function learning level. A fruitful future direction therefore would be to extend our model to allow for the addition of new features or to incorporate attention dynamics.

While the estimation task is sufficient to draw participants' attention to the novel option and make them reflect upon its value in relation to the other options, we do not believe it is necessary to produce the expected effects. There may well be other ways in which people will notice and reflect on novel options without asking them to provide explicit ratings of expected reward and the associated uncertainty. In experimental tasks such as the feature-based bandit task used here, participants make relatively rapid and repeated choices in a simple and highly constrained environment. In daily life, choice sets may be less clearly defined, and a particular choice task will be interspersed with many other tasks. In such situations, people may naturally pay attention to novel options and take time to reflect on the value of all options in the consideration set (Knox et al., 2012). Interestingly, the proportion of function learners was approximately 60% in Experiments 3a-3c, which is substantially higher than the 40% obtained in Experiment 1 and 2. It is likely that the response delay of two seconds in Experiment 3 already pushed participants to reflect more on the task and their approach to it, at least initially. While further increasing the response delay could make participants reflect on all options also on trial 41, it may also have an adverse effect, making people disengage with the task during a prolonged forced holdout period. We leave such fine-tuning of task parameters to future studies.

General discussion

As people sample options in their environment, they face a steady stream of choice dilemmas between novel and tried-

[6] Due to a technical error, the estimation task data on trial 41 was not recorded. We do not report the results of the estimation task from the 70th trial here, because they are less informative.

and-tested options. Traditional models of reinforcement learning do not cope well with such problems, as they lack a mechanism for identifying promising options in a sea of novel possibilities. Nonetheless, people manage to navigate the exploration-exploitation trade-off in realistic and information-rich settings, identifying and choosing options that are not only novel but also good. How is this adaptive feature of human intelligence accomplished?

We have put forward a model that combines functional generalization with uncertainty guidance to describe participants' responses in the face of novelty. We believe that our model can explain parts of this puzzle. The model does not only explain why participants sometimes seek out and sometimes avoid novel options – they generalize their functional knowledge – but it also tells us why they might prefer novel options by default – they are curious about options that they perceive as more uncertain. We used simulations of our model to generate qualitative predictions about people's behavior in a feature-based multi-armed bandit task, contrasting it with a competing model which lacks the function learning component. We tested our predictions in three preregistered experiments. In the first experiment, we found that functional generalization can lead to both seeking out and shunning away from novel options if their features indicate either high or low expected rewards. In the second experiment, we showed that uncertainty guidance can lead to a small but detectable preference for novel and exotic options which are dissimilar to known options compared to "ordinary novel" options that have feature values inside the experienced range. In the third experiment, we further assessed the role of attention and reflection in functional generalization. The results showed that functional generalization to novel options requires participants to pay attention to a novel option and reflect on values of the options.

We found further support for functional generalization and uncertainty guidance by analyzing participants' estimates of expected rewards and their uncertainty about all the options. In particular, confidence ratings in Experiment 1 supported an interesting prediction from our theory, that participants will be confident about their knowledge in consequential, highly rewarding region of the feature space, whereas they would be less knowledgeable about low-rewarding feature values. Interestingly, participants' estimates in Experiment 2 seemed to correspond more closely to those of a similarity-based function learning model than to a rule-based function learning model. When they made predictions for options with features from outside the experienced range, these predictions seemed to revert back to the prior mean, much more than what would be expected if extrapolation relied on a linear function. The resulting underestimation of the reward of the exotic-novel option can explain why, albeit reliable, the uncertainty guidance effect was small. Additionally, the magnitude of uncertainty bonuses (as formalized by the β parameter in the UCB rule) might have been smaller than assumed in the model simulations, which would also reduce the observable effects of uncertainty guidance.

A clear discrepancy between our preregistered hypotheses and the observed behavior in Experiment 1 and 2 was that some of the predicted effects did not occur immediately on the 41st trial in which the novel option was first introduced. Rather, they occurred a trial later, i.e. on the 42nd trial. We investigated this issue further in Experiment 3, using the FMAB conditions from Experiment 1. We obtained weak evidence for the predicted effect when we matched the fading-in time of the new option to the two second period during which participants could not register their choice. Visual effects likely interfered with functional generalization, suggesting that future theories of novelty should explicitly take into account attention dynamics. When the estimation task occurred simultaneously with the introduction of the novel option, and before participants could register their choice, we found strong evidence of the expected effect on the 41st trial. These results indicate that our predictions hold when participants are both aware of the novel option and reflect on the options in the consideration set before making a choice.

How is functional generalization and uncertainty guidance implemented?

Our theory of experiential decision making in information-rich environments purports that people rely on functional generalization and uncertainty guidance. Functional generalization in different environments requires a flexible way to represent and learn functional relations from limited observations. We believe that Gaussian process regression is a useful working model for how people may approach such function learning. GP models can learn a wide variety of functional forms, by using different kernels (e.g., an RBF or linear kernel), and even combining different kernels. Questions such as whether the brain performs computations that correspond to those of a GP regression model, and how a kernel for generalization is chosen and/or learned, are important but beyond the scope of the current contribution. Of interest in Experiment 1 was whether people use functional generalization at all when encountering novel options. We designed our experiment to answer this question, not to contrast a GP-based function learning model to other models of function learning. Similarly, Experiment 2 was concerned with the question whether people are guided by functional uncertainty when they explore novel options. While our model implemented such guidance through the UCB rule, other forms of uncertainty-guided exploration, such as Thompson sampling, would have made qualitatively similar predictions. Experiment 2 was designed to identify functional uncertainty guidance, not to arbitrate between different implementations of it. Prior research has shown that a GP-UCB model describes people's behavior better than other models in a variety of contextual bandit tasks (e.g. Schulz, Konstantinidis, & Speekenbrink, 2018; Wu, Schulz, Garvert, et al., 2018). Having found evidence for both functional generalization and uncertainty guidance, we leave determining the more precise details of these processes to future research.

Redefining novelty

Positioning novelty within our functional generalization and uncertainty guidance framework may provide new insights into the very concept of novelty. Rather than a binary distinction between novel and old options, novelty is a more gradual construct. When are options perceived as more novel? According to our theory, novelty is related to functional uncertainty: when people are more uncertain in generalizing their functional knowledge to new options they experience them as more novel, as compared to when they are less uncertain (the distinction between exotic and ordinary in our paradigm). Because all options were governed by the same feature-reward function, it could be argued that the novel options in our experiments were never really "truly novel". We believe that the same can be said about other studies addressing novelty, where novel options are introduced within the same experimental context as old options. Novelty, in our view, depends on the extent to which prior experience is expected to have a bearing on newly-introduced options. The set-up in our experiments is akin to a new beer appearing alongside familiar ones on the shelve of your supermarket; a new instance of a familiar category.

How would we react to an instance of a new category appearing, or how would we interpret a completely new feature? From a functional generalization perspective, we are likely to find the most similar categories, features, or experiences, and transfer as much knowledge as we can from them (Lucas, Sterling, & Kemp, 2012). We have not addressed how knowledge of a function in one domain may be generalized to form expectations and inform learning new functions in different domains. The question of how learning can be transferred across tasks is currently at the frontier of machine learning research (e.g. Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016; Wang et al., 2016) and constitutes an exciting extension of our framework. Developments in this direction could also be useful for explaining the results of Experiment 3a. When the novel option was introduced with a salient flickering visual effect, the large majority of participants chose it, regardless of its feature values. If people consider novel options to have additional novel features, the old feature-reward function would not hold for the novel option, making transfer of learning a relevant mechanism of how knowledge is transferred to options which only partly share features with other options.

Individual differences and strategy selection

Based on past results using the FMAB paradigm (Stojic, 2016), we expected to find that some participants would not learn the feature-reward function. Indeed, roughly 60% of participants in Experiment 1 and 2, and 40% in Experiment 3 were classified as reward trackers by our functional knowledge task. Not every participant classified as reward tracker might have lacked functional knowledge completely. Due to the coarseness of our preregistered classification procedure, function learners with some but relatively poor functional knowledge were likely misclassified as reward trackers. While imperfect, we believe that our classification procedure allowed us to detect qualitatively different strategies. As predicted, the behavior of the function-learning subgroup corresponded more to the GP-UCB model simulations than the behavior of participants in the FMAB conditions taken as a whole. Moreover, behavior of the reward-tracking subgroup matched the behavior of our control MAB conditions, further supporting our conclusion that different strategies, and not just degrees of functional knowledge, explain our results.

This result leaves us with an important open question: why and how do people opt for either a function-learning or a reward-tracking strategy? One possibility is that such strategy selection reflects stable individual differences, for example, in participants' working memory or IQ. Examining this explanation seriously would require longitudinal studies that additionally examine the relevant traits. However, a similar explanation of people's strategy selection has been insufficient in other domains (Bröder, 2012). Another possibility is that people engage in a cost-benefit arbitration between different strategies, trading off the cognitive costs of applying strategies with their expected benefits. This explanation has been more successful in other domains (Kool, Gershman, & Cushman, 2017; Payne et al., 1993). If the initial choice set is small and stable, functional generalization may not be worth the cognitive effort. Ignoring the feature values whilst trying options may even reduce loss if prior beliefs about the reward function are incorrect (Stojic, 2016). However, if the choice set is large, as in cultural goods markets populated with an immense number of movies or books (Analytis, Stojic, & Moussaïd, 2015; Salganik, Dodds, & Watts, 2006), a function-learning strategy would work much better. In fact, in such environments people would likely expect novel options to constantly appear and correspondingly expect a need to generalize. How would people resolve this trade-off? They might learn which strategy has the best cost-benefit trade-off (Lieder & Griffiths, 2017; Rieskamp & Otto, 2006; Stojic, Olsson, & Speekenbrink, 2016) or arbitrate between strategies based on the relative uncertainty with which these strategies predict rewards (Daw, Niv, & Dayan, 2005). Regardless of the exact mechanism, identifying why and when some people employ a function-learning strategy whilst others a reward-tracking strategy is a valuable line of future research.

While we have focused on a qualitative distinction between function learners and reward trackers, it is likely that function learners themselves differed in their ability to learn the reward function. Such differences can be captured within our GP-UCB model. Differences in perceptual and memory noise (i.e. perceiving differences between visually presented feature values and recalling experienced rewards) can be modeled as differences in the noise variance σ_ϵ^2, while differences in prior assumptions can be reflected by the choice of kernel (e.g., whether a linear or RBF kernel, as well as parameters of particular kernels, such as the length scale λ of the RBF kernel), or by adapting the initial mean function (e.g., a positive linear initial mean function). Identifying such individual differences may provide more insight into why not everyone who employs a function learning strategy has the same level of functional knowledge. While our theoretical framework is sufficiently rich to characterize individual differences, we designed our experiments to be mainly sensitive to differences between function learners and reward trackers, not between more subtle differences within the group of function learners. Assessing people's prior beliefs about feature-reward functions, and differences in their learning, will require studies tailored to these goals.

Task horizon effects

People's exploration is affected by the task horizon: people normally decrease the amount of exploration with the number of choices left. This is predicted by rational models (R. C. Wilson et al., 2014). Currently, the GP-UCB model does not incorporate such a dynamic exploration policy. It is straightforward to include it in a heuristic manner, by decreasing the exploration parameter (β) over time. Future studies could explore this modification. Another way, closer to an optimal solution, would be to combine our model with recently developed approximate approaches to Bayesian planning under model uncertainty (Gonzalez, Osborne, & Lawrence, 2016; Guez, Silver, & Dayan, 2013). Planning optimally in non-trivial tasks is notoriously difficult and approximations are generally necessary. Stochastic planning by Monte Carlo tree search (Browne et al., 2012; Guez et al., 2013) has firmer normative grounds than simply decreasing an exploration parameter over time. Notably, Krusche, Schulz, Guez, and Speekenbrink (2018) and van Opheusden, Galbiati, Bnaya, Li, and Ma (2017) found empirical evidence for such strategies in challenging decision making tasks.

Function learning in the wild

Our work goes beyond traditional function learning paradigms, and introduces a new—yet commonly encountered—setting for function learning, where people need to balance acquiring new information with choosing rewarding options. In traditional function learning paradigms, people are passive information gatherers, learning from stimuli selected by the experimenters. As such, it is unclear how well extant findings generalize to real-life settings where people choose the stimuli (options) to learn about, whilst simultaneously being concerned with how those stimuli serve other goals (i.e, obtaining rewards). Research on active forms of information gathering has mostly focused on purely exploratory settings where the goal is solely to maximize information (Bramley, Lagnado, & Speekenbrink, 2015; Nelson, 2005; Nelson, McKenzie, Cottrell, & Sejnowski, 2010), or where information acquisition and utility maximization are cast as competing goals (Markant & Gureckis, 2012; Meder & Nelson, 2012). In our reinforcement learning paradigm, function learning supports utility maximization, and exploration and maximization are not competing, but rather compatible goals (see Rich & Gureckis, 2018; Zhang & Angela, 2013, for similar arguments). Our results indicate that in such a setting, people are motivated by both short-term utility gains and the long-term consequences of information gains, instead of focusing exclusively on one or the other. As a result, people gain more experience in consequential regions where feature values are likely to be rewarding. Accordingly, people are confident about their functional knowledge in that region of the feature space, whereas they remain less knowledgeable about low-rewarding feature values. Our results suggest that functional knowledge in the wild is likely to be skewed in systematic ways by choices and goals, potentially resulting in phenomena such as polarization of beliefs (Bénabou & Tirole, 2016) or illusory correlations (Denrell & Le Mens, 2011; Hogarth, Lejarraga, & Soyer, 2015).

Understanding consumer behavior

Many consumer choice settings match well with the problem we have studied—the choice set is commonly large, products have multiple features and consumers make purchases repeatedly. As of the late 1980s, marketing scholars have developed formal learning models to capture consumer behavior in such settings (Roberts & Urban, 1988). Typically, people are assumed to have initial expectations about the quality of products that they update on the basis of word of mouth or their direct experiences with the product (e.g. Ching et al., 2013). Learning models in marketing can capture effects, such as brand loyalty, that are hard to accommodate for their non-learning counterparts (e.g. Guadagni & Little, 1983). However, they tend to assume either too much or too little of human cognition. For instance, they lack a clear mechanism for integrating cues (Lin, Zhang, & Hauser, 2014) or they purport that people plan deep in the future (Erdem & Keane, 1996), an assumption that has received little empirical support in behavioral studies (e.g. Gabaix, Laibson, Moloche, & Weinberg, 2006).

Going beyond learning models in marketing, the GP-UCB model provides a psychologically founded account of how people integrate different features, and use functional generalization and uncertainty to guide their decisions in consumer choice settings. Further, the model can be adequately specified against real world data. A recent study by Schulz, Bhui, et al. (2019) provides evidence that the model can capture consumer behavior when people choose repeatedly among different options in the wild. The authors analyzed a large real world data set of customers' online food delivery orders and showed that the GP-UCB model describes well how people allocate choices among numerous restaurants.

The GP-UCB model and the results from our experiments can be used to predict when a consumer will explore a new product (Hirschman, 1980; Riefer, Prior, Blair, Pavey, & Love, 2017). Uncertainty may be a crucial factor to tempt customers to try a novel product. Companies could aim at building products that strike a good balance between evoking evaluations of high quality and enticing consumers due to their original design. What is more, our modeling account can capture other key empirical phenomena in consumer choice such as variety seeking and brand loyalty (Kahn, 1995). In fact, these two phenomena may be accounted for by a single learning process—that is, how people use features to form expectations and uncertainty to balance the exploration-exploitation trade-off. This is a simpler and more elegant explanation than extant formal approaches that often assume that people directly derive utility from seeking a variety of options (McAlister & Pessemier, 1982; Ratner, Kahn, & Kahneman, 1999).

Finally, our results in Experiment 3 suggest that attention likely plays an important role over and above functional generalization and uncertainty guidance. Experiment 3a, in particular, suggests that increasing the salience of an option directly increases the probability of testing it, regardless of its' expected quality. Thus, it may have captured the effect of aggressive advertising of novel options that people experience in every day life. Our GP-UCB model currently assumes perfect attention and cannot explain such attentional effects. Incorporating attention dynamics in the model would likely lead to substantial increase in explanatory power of the model. This theoretical development can be guided by the recent work focusing on interplay between attention and reinforcement learning (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Niv et al., 2015; Radulescu, Niv, & Ballard, 2019; Stojic, Orquin, Dayan, Dolan, & Speekenbrink, 2020).

Concluding remarks

In summary, we believe that our theory offers a powerful and expressive account of human behavior in the face of novelty. The core claim of our theory is that people use functional generalization and are guided by uncertainty when confronted with novel options. Our results do not make specific claims about the precise implementation of these mechanisms (i.e. an RBF kernel combined with UCB sampling). Instead, they strongly suggest that an account of people's behavior requires both a model of functional generalization and an exploration strategy that attempts to reduce uncertainty about those generalizations. Beyond novelty, integration of function learning and decision making allows revisiting familiar problems from a new perspective and opens up new avenues of research. Studying how people use generalization and uncertainty to guide their choices in complex decision-making tasks will continue to revise our picture of human intelligence. For this, we need to keep exploring.

Context of the research

This work evolved within a broader research program that aims to understand how people use generalization to efficiently explore their environments in the search for rewards (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojic et al., 2015; Wu, Schulz, Speekenbrink, et al., 2018). In our earlier work, we have shown that participants can apply function learning to guide their search in contextual bandits (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojic et al., 2015), in bandits with spatial correlations between rewards (Wu, Schulz, Speekenbrink, et al., 2018), and in bandits with no explicit relation between features and rewards (Schulz, Franklin, & Gershman, 2018; Stojic, 2016). We thought that the same approach can be used to improve our understanding of novelty. In contrast to our previous research, rather than relying on model fitting, in this study we aimed to design our experiments to allow for a direct test of key theoretical predictions. This was a first "experiment" with preregistrations for all of us. We found the process very valuable, even if we originally did not consider all possibilities such as the delayed notice of the novel option. Preregistering our hypotheses forced us to think through the modeling, predictions and experimental design in greater detail. We also found presenting our to-be-preregistered ideas at conferences rewarding, allowing us to obtain advice for improving the study, even before data collection commenced. In future work, we aim to extend our model to assess the neural correlates of generalization and uncertainty-guided exploration, test how people track their uncertainty in other domains that require functional knowledge (cf. Stojic, Hrvoje and Eldar, Eran and Bassam, Hassan and Dayan, Peter and Dolan, Raymond, 2018), as well as extend our model further to real world decision making such as consumer behavior.

Acknowledgments

We would like to thank Robin Hogarth, Gael Le Mens, Peter Dayan, Thorbjorn Knudsen, Thorsten Pachur and Thorsten Joachims for their feedback. H.S. was supported by the Ministerio de Educación, Cultura y Deporte, grant number FPU12/05859. E.S. is supported by the Harvard Data Science Initiative. This research was supported in part through NSF Award IIS-1513692.

References

Acuna, D., & Schrater, P. R. (2009). Structure learning in human sequential decision-making. In *Advances in Neural Information Processing Systems* (pp. 1–8).

Analytis, P. P., Kothiyal, A., & Katsikopoulos, K. V. (2014). Multiattribute utility models as cognitive search engines. *Judgment and Decision Making*, *9*, 403–419.

Analytis, P. P., Stojic, H., & Moussaïd, M. (2015). The collective dynamics of sequential search in markets for cultural products. *Santa Fe Institute Working Paper*.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*, 235–256.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*, 141–64.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics*, *8*, 279–286.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, *30*, 79–101.

Blanchard, T. C., Hayden, B. Y., & Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron*, *85*, 602–614.

Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *2019*, niz004. doi: 10.1093/nc/niz004

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 55–63).

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 708.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*, 1–27.

Bröder, A. (2012). The quest for take the best - Insights and outlooks from experimental research. In P. M. Todd, G. Gigerenzer, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 216–240). New York, NY, US: Oxford University Press.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., … Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, *4*, 1–43.

Bunzeck, N., & Düzel, E. (2006). Absolute Coding of Stimulus Novelty in the Human Substantia Nigra/VTA. *Neuron*, *51*, 369–379. doi: 10.1016/j.neuron.2006.06.021

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts and categories. studies in cognition.* (pp. 408–437). Cambridge, MA, US: MIT Press.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. doi: 10.18637/jss.v080.i01

Carpenter, A. C., & Schacter, D. L. (2016). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*, 1–144.

Ching, A. T., Erdem, T., & Keane, M. P. (2013). Learning models: An assessment of progress, challenges, and new developments. *Marketing Science*, *32*, 913–938.

Cowan, P. (1976). The new object reaction of rattus rattus l.: the relative importance of various cues. *Behavioral Biology*, *16*, 31–44.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One*, *8*, e57410. doi: 10.1371/journal.pone.0057410

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711. doi: 10.1038/nn1560

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

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986.

Denrell, J., & Le Mens, G. (2011). Seeking positive experiences can produce illusory correlations. *Cognition*, *119*, 313–324.

Erdem, T., & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science*, *15*, 1–20.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*, 659–676.

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*, 0002.

Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, *96*, 1043–1068.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science*, *7*, 391–415.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*, 148–164.

Gonzalez, J., Osborne, M., & Lawrence, N. (2016). Glasses: Relieving the myopia of bayesian optimisation. In A. Gretton & C. C. Robert (Eds.), *Proceedings of the 19th international conference on artificial intelligence and statistics* (Vol. 51, pp. 790–799). Cadiz, Spain: PMLR.

Guadagni, P. M., & Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science*, *2*, 203–238.

Guez, A., Silver, D., & Dayan, P. (2013). Scalable and efficient bayes-adaptive reinforcement learning based on Monte-Carlo tree search. *Journal of Artificial Intelligence Research*, *48*, 841–883. doi: 10.1613/jair.4117

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 1–14. doi: 10.3758/s13428-015-0642-8

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, *62*, 255–262.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*, 245–258.

Hirschman, E. C. (1980). Innovativeness, novelty seeking, and consumer creativity. *Journal of Consumer Research*, *7*, 283–295.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1193–1217. doi: 10.1037/xlm0000241

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, *24*, 379–385.

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., & Wagenmakers, E.-J. (2017). Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods*, *49*, 638–652. doi: 10.3758/s13428-016-0739-8

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 924–941. doi: 10.1037/0278-7393.29.5.924

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*, 133–156.

Kahn, B. E. (1995). Consumer variety-seeking among goods and services: An integrative review. *Journal of Retailing and Consumer Services*, *2*, 139–148.

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*, 549–559. doi: 10.1016/S0893-6080(02)00048-5

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*, 1072–1099.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Jounal of the American Statistical Association*, *90*, 773–795.

Keller, K. L. (2002). Branding and brand equity. In *Handbook of marketing* (p. 151-178). Thousand Oaks, CA: Sage Publications.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *2:398*, 1–12.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 811–836.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*, 1321–1333.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*, 1292–1298. doi: 10.1038/nn.2635

Krause, A., Singh, A., & Guestrin, C. (2008). Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, *9*, 235–284.

Krusche, M. J., Schulz, E., Guez, A., & Speekenbrink, M. (2018). Adaptive planning in human search. *bioRxiv*, 268938.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. doi: 10.1037/0033-295X.99.1.22

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Langford, J., & Zhang, T. (2008). The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems* (Vol. 20, pp. 817–824). Curran Associates, Inc.

Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*, 1–12. doi: 10.3758/s13428-014-0458-y

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*, 379–392.

Le Mens, G., Kareev, Y., & Avrahami, J. (2016). The evaluative advantage of novel alternatives an information-sampling account. *Psychological Science*, *27*, 161–168.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*, 451–463.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 661–670). ACM Press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, *103*, 410–423. doi: 10.1198/016214507000001337

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*, 762–794.

Lin, S., Zhang, J., & Hauser, J. R. (2014). Learning from experience, simply. *Marketing Science*, *34*, 1–19.

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, *110*, 6139–6144. doi: 10.1073/pnas.1217854110

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*, 1193–1215. doi: 10.3758/s13423-015-0808-5

Lucas, C. G., Griffiths, T. L., Xu, F., & Fawcett, C. (2009). A rational model of preference learning and choice prediction by children. In *Advances in Neural Information Processing Systems* (pp. 985–992).

Lucas, C. G., Sterling, D., & Kemp, C. (2012). Superspace extrapolation reveals inductive biases in function learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

Mahajan, V., Muller, E., & Srivastava, R. K. (1990). Determination of adopter categories by using innovation diffusion models. *Journal of Marketing Research*, 37–50.

Markant, D., & Gureckis, T. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

Markant, D., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*, 100–120.

McAlister, L., & Pessemier, E. (1982). Variety seeking behavior: An interdisciplinary review. *Journal of Consumer research*, *9*, 311–322.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*, 119–148.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*, 406–419. doi: 10.1037/a0024377

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software manual]. Retrieved from `https://cran.r-project.org/package=BayesFactor` (R package version 0.9.12-2)

Neal, R. M. (1996). *Bayesian learning for neural networks*. Springer Verlag.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979-999.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969.

Nissen, H. W. (1930). A study of exploratory behavior in the white rat by means of the obstruction method. *The Pedagogical Seminary and Journal of Genetic Psychology*, *37*, 361–376.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*, 8145–8157. doi: 10.1523/JNEUROSCI.2978-14.2015

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114. doi: 10.1037/0278-7393.10.1.104

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–61.

Nunnally, J. C., & Lemond, L. C. (1974). Exploratory behavior and human development. *Advances in Child Development and Behavior*, *8*, 59–109.

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*, 8096. doi: 10.1038/ncomms9096

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*, 184–188. doi: 10.1177/0963721414531598

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, *7*, e1001048.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: the role of structure and attention. *Trends in Cognitive Sciences*, *23*(4), 278–292. doi: 10.1016/j.tics.2019.01.010

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ratner, R. K., Kahn, B. E., & Kahneman, D. (1999). Choosing less-preferred experiences for the sake of variety. *Journal of Consumer Research*, *26*, 1–15.

Reichel, C. M., & Bevins, R. A. (2008). Competition between the conditioned rewarding effects of cocaine and novelty. *Behavioral Neuroscience*, *122*, 140–150.

Rich, A. S., & Gureckis, T. M. (2018). Exploratory choice reflects the future value of information. *Decision*, *5*, 177.

Riefer, P. S., Prior, R., Blair, N., Pavey, G., & Love, B. C. (2017). Coherency-maximizing exploration in the supermarket. *Nature human behaviour*, *1*, 0017.

Rieskamp, J., & Otto, P. E. (2006). SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.

Rigoli, F., Friston, K. J., & Dolan, R. J. (2016). Neural processes mediating contextual influences on human choice behaviour. *Nature Communications*, *7*, 12416. doi: 10.1038/ncomms12416

Roberts, J. H., & Urban, G. L. (1988). Modeling multiattribute utility, risk, and belief dynamics for new consumer durable brand choice. *Management Science*, *34*, 167–185.

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Re-*

search, *47*, 877–903. doi: 10.1080/00273171.2012.734737

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi: 10.3758/PBR.16.2.225

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*, 854–856.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning* (pp. 1842–1850).

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, 13903-13908. doi: 10.1073/pnas.1821028116

Schulz, E., Franklin, N. T., & Gershman, S. J. (2018). Finding structure in multi-armed bandits. *BioRxiv*, 432534.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 927–943.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological science*, *30*, 1561–1572.

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*, 1317.

Smith, S. M., & Krajbich, I. (2019). Gaze amplifies value in decision making. *Psychological Science*, *30*, 116–128.

Speekenbrink, M., Channon, S., & Shanks, D. R. (2008). Learning strategies in amnesia. *Neuroscience and Biobehavioral Reviews*, *32*, 292–310. doi: 10.1016/j.neubiorev.2007.07.005

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, *7*, 351–367. doi: 10.1111/tops.12145

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychological: General*, *139*, 266–298. doi: 10.1037/a0018620

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2012, May). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, *58*, 3250-3265. doi: 10.1109/TIT.2011.2182033

Stan Development Team. (2018). *RStan: the R interface to Stan.* Retrieved from http://mc-stan.org/ (R package version 2.17.3)

Steenkamp, J.-B. E. M., & Gielens, K. (2003). Consumer and Market Drivers of the Trial Probability of New Consumer Packaged Goods. *Journal of Consumer Research*, *30*, 368–384. doi: 10.1086/378615

Stojic, H. (2016). *Strategy selection and function learning in decision making* (Doctoral dissertation, Universitat Pompeu Fabra). Retrieved from http://hdl.handle.net/10803/400136

Stojic, H., Analytis, P. P., & Speekenbrink, M. (2015). Human behavior in contextual multi-armed bandit problems. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society* (pp. 2290–2295).

Stojic, H., Olsson, H., & Speekenbrink, M. (2016). Not everything looks like a nail: Learning to select appropriate decision strategies in multiple environments. *PsyArXiv*.

Stojic, H., Orquin, J., Dayan, P., Dolan, R., & Speekenbrink, M. (2020). Uncertainty in learning, choice and visual fixation. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1911348117

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2018a). *It's new, but is it good? how generalization and uncertainty guide the exploration of novel options.* PsyArXiv. Retrieved from psyarxiv.com/p6zev doi: 10.31234/osf.io/p6zev

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2018b). *Preregistration for "It's new, but is it good? How generalization and uncertainty guide the exploration of novel options".* Open Science Framework. Retrieved from https://osf.io/upj76 doi: 10.17605/OSF.IO/UPJ76

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2018c). *Project files for "It's new, but is it good? How generalization and uncertainty guide the exploration of novel options".* Open Science Framework. Retrieved from https://osf.io/c8u9t/

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2019a). *Preregistration for Experiment 3a in "It's new, but is it good? How generalization and uncertainty guide the exploration of novel options".* Open Science Framework. Retrieved from https://osf.io/h5uqr

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2019b). *Preregistration for Experiment 3b in "It's new, but is it good? How generalization and uncertainty guide the exploration of novel options".* Open Science Framework. Retrieved from https://osf.io/37ayn

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2019c). *Preregistration for Experiment 3c in "It's new, but is it good? How generalization and uncertainty guide the exploration of novel options".* Open Science Framework. Retrieved from https://osf.io/tg5kc

Stojic, Hrvoje and Eldar, Eran and Bassam, Hassan and Dayan, Peter and Dolan, Raymond. (2018). Are you sure about that? On the origins of confidence in concept learning. In *Proceedings of the Cognitive Computational Neuroscience Conference* (pp. 55–63).

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA, US: MIT Press.

Teodorescu, K., & Erev, I. (2014). On the decision to explore new alternatives: The coexistence of under-and over-exploration. *Journal of Behavioral Decision Making*, *27*, 109–123.

Thompson, W. R. (1933). On the Likelihood that One Unknown

Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, *25*, 285–294. doi: 10.2307/2332286

Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 39).

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, *137*, 73–96.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 627–633. doi: 10.1177/1745691612463078

Wang, J., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J., Munos, R., … Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Whittle, P. (1980). Multi-Armed Bandits and the Gittins Index. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*, 143–149.

Wilson, A. G., Dann, C., Lucas, C. G., & Xing, E. P. (2015). The Human Kernel. In *Advances in Neural Information Processing Systems* (pp. 2854–2862).

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*, 2074–2081. doi: 10.1037/a0038199

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). Connecting conceptual and spatial search via a model of generalization. *bioRxiv*, 258665.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*. doi: 10.1038/s41562-018-0467-4

Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*, 224–228.

Zhang, S., & Angela, J. Y. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (pp. 2607–2615).

Appendix A

Functional knowledge task

Participants in the FMAB conditions completed an additional functional knowledge task (Figure 3). Here we provide details on constructing the stimuli for that task.

We designed the choice triplets such that there was always a best, a medium, and a worst option. With perfect knowledge of the function, participants should be able to identify the options as such. There were three types of items – 5 "easy", 10 "difficult", and 10 special "weight comparison" items with choice triplets designed to detect whether people have learned which feature has a greater impact on rewards. Denoting the best, medium and worst option as x, y and z, the feature values were generated as follows:

- In the easy triplets, the best option had lower (better) feature values on both dimensions than the medium option, and the medium option had lower (better) feature values on both dimensions than the worst option. To construct these triplets, we first randomly drew the feature values for the best option. After this, the intervals for sampling the feature values of the medium option were generated. We constructed the feature values of the worst option in analogous way.

  1. $x_1 \sim U(0.2, 0.6),\ x_2 \sim U(0.2, 0.6)$
  2. $y_1 \sim U(x_1 + 0.05, 0.7),\ y_2 \sim U(x_2 + 0.05, 0.7)$
  3. $z_1 \sim U(y_1 + 0.05, 0.8),\ z_2 \sim U(y_2 + 0.05, 0.8)$

- We constructed the "difficult" triplets in a similar manner, but the intervals for sampling the features of the medium and worst option were closer to the features of the best option, making the medium and worst option more similar to the best option than in the easy type.

  1. $x_1 \sim U(0.2, 0.6)$
  2. $y_1 \sim U(x_1 + 0.05, 0.7),\ y_2 \sim U(0.2, 0.6)$
  3. $x_2 \sim U(y_2 - 0.05, \min(0.7, y_2 + \frac{w_1}{w_2}(y_1 - x_1)))$
  4. $z_1 \sim U(\max(x_1, y_1) + 0.05, 0.8)$
  5. $z_2 \sim U(\max(x_2, y_2) + 0.05, 0.8)$

- The "weight comparison" items consisted of a best option with a small value of the feature with the largest weight and a large value of the other feature. The medium option had exactly the opposite pattern, thereby creating a diagnostic pair for detecting whether people have learned which feature is more predictive. The worst option had two large feature values.

  1. $x_1 \sim U(0.25, 0.35),\ x_2 \sim U(0.7, 0.8)$
  2. $y_1 \sim U(0.7, 0.8),\ y_2 \sim U(0.25, 0.35)$
  3. $z_1 \sim U(0.7, 0.8),\ z_2 \sim U(0.7, 0.8)$

To achieve good performance in this task, participants had to use their functional knowledge acquired in the bandit task. Our preregistered classification procedure uses participants' mean performance in the task to classify participants who achieved better-than-chance performance as function learners, and participants who performed at chance-level or worse as reward trackers. Our classification procedure was to classify participants as function learners if the Bayes factor comparing a model in which the true mean rank is better (lower) than chance performance (mean rank 2) to a model in which this equals chance performance provided "substantial" evidence (BF $\geq$ 10) for the first model.

We performed simulations to examine the sensitivity of our classification procedure. We simulated function learners starting from perfect knowledge, performing optimally in the functional knowledge task, to progressively worse performing function learners until random performance, the performance level of reward trackers. More precisely, for all simulated participants, we drew predicted reward values for each option from a Normal distribution centered on the actual expected reward of each option (as derived from the feature-reward function). To vary the level of functional knowledge, we varied the standard deviation (from 0 to 220 in steps of 1) of these distributions. Increasing the level of noise (standard deviation) from which predictions were drawn results in progressively worse performance. We simulated 100 participants for each knowledge level, applying our classification procedure to the choices of each simulated participant. Figure A1 shows the proportion of simulated participants for which the Bayes factor was greater than 10 (our classification criterion), indicating the probability of classifying a participant as function learner (denoted as P(function learner)) as a function of their score in the functional knowledge test. These simulation results show that our classification procedure is reasonably likely to correctly classify participants with moderate or better functional knowledge as function learners, while those with poorer knowledge are likely to be miss-classified as reward trackers. Note also that if reward trackers are choosing randomly, the chance of falsely classifying them as function learners was very low (0.65% for 10000 simulated random choices), as with random choices it is unlikely to reach a good enough mean rank to be classified as a function learner.

Data analysis

We followed the recommendations of Wagenmakers, Wetzels, Borsboom, van der Maas, and Kievit (2012) with regards to data collection and analysis, and relied on Bayesian statistics throughout.

Recorded variables. For the sake of full transparency, we recorded the following variables in our experiments: (1) participants' choices and response times in the bandit task and functional knowledge task; (2) estimates, confidence ratings and response times in the estimation task; (3) age, gender, and whether they had noticed the appearance of the novel option in a questionnaire at the end of the experiment (Experiment 3b and 3c included additional exploratory questions probing participants' knowledge of the reward function).

Sampling plan. We planned to collect a minimum of 60 participants in each of the four between-subject condition. Thereafter, we evaluated the Bayes factor of the tests of our main hypotheses concerning the proportion of choices allocated to the novel option on 41st trial. We proceeded with data collection iteratively, collecting batches of 5 additional participants in each FMAB condition, stopping as soon as we reached "strong" evidence (and continuing data collection otherwise). We defined "strong" evidence as a Bayes factor of 10 or larger in favor of either the null or alternative hypothesis (Jeffreys, 1961). We derived the minimum by performing the main hypothesis tests on simulated data with predicted differences. Given our main hypothesis, we increased the number of participants only in the FMAB conditions. We also planned to stop the experiments in case we run out of funds, which corresponded to a maximum of approximately 750 participants.

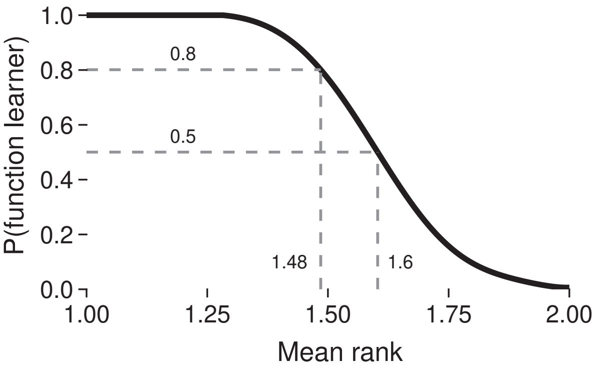


Figure A1. Results of simulating participants with different levels of functional knowledge (expressed here through their mean choice rank) and assessing the sensitivity of our preregistered procedure for classifying participants as function learners (denoted as P(function learner)). Results show that our classification procedure is reasonably likely to correctly classify participants with moderate and better knowledge, while those with poorer knowledge are likely to be miss-classified as reward trackers.

Statistical tests. We use Bayes factors to quantify the relative evidence the data provides in favor of the null (H_0) or the alternative hypothesis (H_1). The Bayes factor quantifies the probability of the data under H_0 relative to the probability of the data under H_1 (e.g. Kass & Raftery, 1995). We denote such a Bayes factor as BF_{01}. For example, a BF_{01} of 10 indicates that the data are 10 times more likely under the H_0 than under the H_1. When comparing the H_0 hypothesis relative to the H_1, we express the evidence as BF_{10}. We conducted all tests by using the `BayesFactor` package implemented in R (Morey & Rouder, 2015; R Core Team, 2016).

For hypotheses concerning participants' choices in a single trial, the dependent variable was the proportion of participants in a condition who choose the novel option, while the independent variable was the experimental condition. We used a version of the contingency table Bayes factor test of Jamil et al. (2017), with an independent multinomial sampling assumption and a default "weak" Dirichlet prior (a = 1

Morey & Rouder, 2015).

We used a Bayesian binomial model to estimate probability that participants in a condition would choose a novel option over the course of multiple trials – from trial 41 to 70 in the FMAB conditions in Experiment 1 and from trial 41 to 55 in Experiment 2. It is a hierarchical model that treats participants as members of a group, taking into account the group's probability distribution when estimating individual probabilities (e.g. Kruschke, 2014). This leads to more realistic posterior distribution of the group-wise probabilities we are interested in. We used a non-centered probit parameterization which facilitates Markov chain Monte Carlo (MCMC) sampling when there are small number of observations per participant (Betancourt & Girolami, 2015). We defined the priors of group-level means and standard deviations based on our model simulation results. In Experiment 1 for the FMAB conditions (and function learner subgroups), we assigned priors based on simulations for trials 41 to 70, $\mu \sim N(-2.19, 1)$ (corresponding to a mean probability of $p = 0.01$) for the low value and $\mu \sim N(2.21, 1)$ ($p = 0.99$) for the high value condition. For the MAB conditions, the same priors were $\mu \sim N(-1.83, 1)$ ($p = 0.03$) for the low value and $\mu \sim N(2.75, 1)$ ($p = 0.99$) for the high value condition. In Experiment 2, for the FMAB conditions (and function learner subgroups) priors were based on our simulation results for trials 41 to 46 where models showed a difference, $\mu \sim N(-0.38, 1)$ ($p = 0.35$) for the exotic-novel and $\mu \sim N(-0.61, 1)$ ($p = 0.27$) for the ordinary-novel condition. For the MAB conditions the priors were $\mu \sim N(0.20, 1)$ ($p = 0.58$) for the exotic-novel and $\mu \sim N(0.12, 1)$ ($p = 0.55$) for the ordinary-novel condition. We used the same half-cauchy prior for the group-level standard deviations in all conditions, $\sigma \sim C(0, 1)$. We estimated the model with the NUTS-MCMC algorithm implemented in Stan (Stan Development Team, 2018). We initialized four chains with randomly generated starting values and collected 60000 samples for each chain, after discarding the first 40000 samples as burn-in. We confirmed that all chains had successfully converged by visually inspecting them and examining the $\hat{R}$ statistic. We also confirmed that we correctly implemented the model with parameter recovery studies on simulated data.

For hypotheses related to the estimation task, we used the Bayesian t-test for independent samples proposed by Morey and Rouder (2011); Rouder et al. (2009), with the Jeffreys-Zellner-Siow prior and scale set to $\sqrt{2}/2$. We truncated the prior above or below 0 for directional hypotheses (our default H_1 hypotheses), and used a symmetric prior for non-directional hypotheses (explicitly indicated when used). We used the same default one-sided t-test to classify participants as "function learners" or "reward trackers" based on their performance in the functional knowledge task. We compared the mean rank of participants' choices (rank 1 being the best and 3 being the worst) across all 25 choices in the task to the mean rank of a person choosing fully at random, which would equal a rank of 2. The null hypothesis was that there was no difference, while the alternative hypothesis was that the mean rank was lower than 2. If there was strong evidence ($BF_{10} > 10$) that a participant's mean rank was below 2, we classified the participant as a function learner, and as a reward tracker otherwise.

Further sanity-check hypotheses involved testing the relationship between the number of times an option was chosen and the accuracy of participants' estimates and their confidence. Here, we used linear regression and computed the Bayes factor for the model with a single predictor (the number of times the option had been chosen) against an intercept-only model, again with a Jeffreys-Zellner-Siow prior and scale set to $\sqrt{2}/2$ (Liang, Paulo, Molina, Clyde, & Berger, 2008; Rouder & Morey, 2012). We used the same approach to test hypotheses about a dependence between choice performance and the number of failed attention checks.

Appendix B

Attention check analysis

Participants had to complete a simple attention check after reading instructions and before they could proceed to the bandit task. They had to correctly answer all four questions in the attention check to proceed, otherwise they were returned to the beginning of instructions. Here we examined if participants that needed more attempts to pass the attention check performed worse in the bandit task. There was moderate evidence that failing the attention questions is correlated with performance in the FMAB conditions in Experiments 1 and 2. We assessed this by regressing the number of failed attention checks onto the sum of earned points until trial 40 ($BF_{10} = 3.10$, with an intercept-only model as H_0). We found no evidence of such an effect in the MAB conditions ($BF_{10} = 0.27$).

Analysis of the post-experiment questionnaire

In the questionnaire at the end of the experiment we asked "Did you notice that a new option appeared in the 41st trial, just before the estimation task interrupted the game?". For Experiment 1 and 2, we explored the difference in choosing the novel option on trial 41 between those participants who indicated they had noticed noticed the novel option on that trial (29% in Experiment 1 and 27% in Experiment 2) and those who indicated they had not noticed it. In Experiment 1 there was no evidence for a difference between these groups ($BF_{10} = 0.12$), while there was strong evidence for a difference in Experiment 2 ($BF_{10} = 1.8 \times 10^4$).

Since our hypotheses about choosing the novel option 41st trial in Experiment 1 and 2 potentially failed because participants failed to notice the novel option, we also examined the choices only of participants who indicated at the end of the experiment that they had noticed the novel option on 41st trial. Focusing only on these participants, in Experiment 1 we found no evidence for a difference in choice proportions on 41st trial between the FMAB high value condition and FMAB low value condition ($BF_{10} = 0.28$). Similarly, in Experiment 2 we found no evidence for a difference in choice proportions on 41st trial between the FMAB exotic-novel and FMAB ordinary-novel condition ($BF_{10} = 0.22$).

Since we did not preregister any predictions for this measure and because it was likely hard to precisely remember when one noticed the novel option, we did not analyze this response further.

Functional knowledge task performance and choosing the novel option

In the main text, we used a binary classification of function learners and reward trackers and analyzed the behavior of these two groups separately where needed. An alternative analysis is to correlate performance on the functional knowledge task with outcomes such as choosing the novel option. We chose not to focus on this alternative analysis for various reasons. For example, we aimed to describe qualitative differences between people employing a function learning and people employing a reward tracking strategy and preregistered tests to compare the behavior of these groups. The functional knowledge task was designed with this goal in mind and not to reliably detect small differences in participants' functional knowledge.

Nevertheless, we conducted an analysis to assess the impact of functional knowledge in a more fine-grained manner. In this analysis, we focused only on participants in the FMAB conditions, and we estimated a Bayesian logistic model (Bürkner, 2017), where we regressed mean choice rank from the functional knowledge task onto a binary variable denoting whether participants chose the novel option in trial 42. We used fairly uninformative Normal priors on both the intercept and slope ($N(0, 10)$) and estimated the model separately for each FMAB condition in Experiment 1 and 2. We examined the results of both models after checking that there were no convergence issues in the MCMC chains (four chains in total, 10000 burn-in samples, 20000 samples per chain).

For Experiment 1, in the FMAB high-value condition, there clearly was an effect of functional knowledge on choosing the novel option. The slope of the mean rank was negative, indicating a larger preference for the novel option with more knowledge (lower rank), with a posterior mean of -1.89 and 95% HPDI $[-3.10, -0.77]$. In the FMAB low-value condition, the slope was not clearly different from 0, with a posterior mean of -0.96 and 95% HPDI $[-2.93, 0.88]$. The results for Experiment 2 were similar. In the exotic-novel condition, preference for the novel option was clearly stronger for participants with more functional knowledge. The posterior mean of the slope of mean choice rank was -1.41 and 95% HPDI $[-2.34, -0.53]$. In the ordinary-novel condition, the slope was not clearly different from 0, with a posterior mean of -0.67 and 95% HPDI $[-1.94, 0.51]$.

These results were mostly consistent with our a priori model simulations. In the low value condition, we cannot reliably predict an effect of functional knowledge on avoiding the novel option, as even coarse functional knowledge would allow identifying the novel option as a poor choice. In the high value condition, determining that the novel option is better than the best old option requires much more precise functional knowledge. Predictions of the effect of functional knowledge on choosing the novel option in Experiment 2 are less clear. People with perfect functional knowledge would be able to designate the novel option as worse than the best old option. For obvious reasons, the uncertainty guidance effect can be predicted only for people who are at least somewhat uncertain about the feature-reward function. To what extent small differences in functional knowledge are related to small differences in people's ability to assess a reasonable level of uncertainty in generalization of that functional knowledge is also not immediately obvious. For that reason, we report these analyses for the interested readers here, but place more emphasis on our preregistered binary classification, as predictions are more straightforward when comparing function learners to reward trackers which are assumed to have no functional knowledge at all.

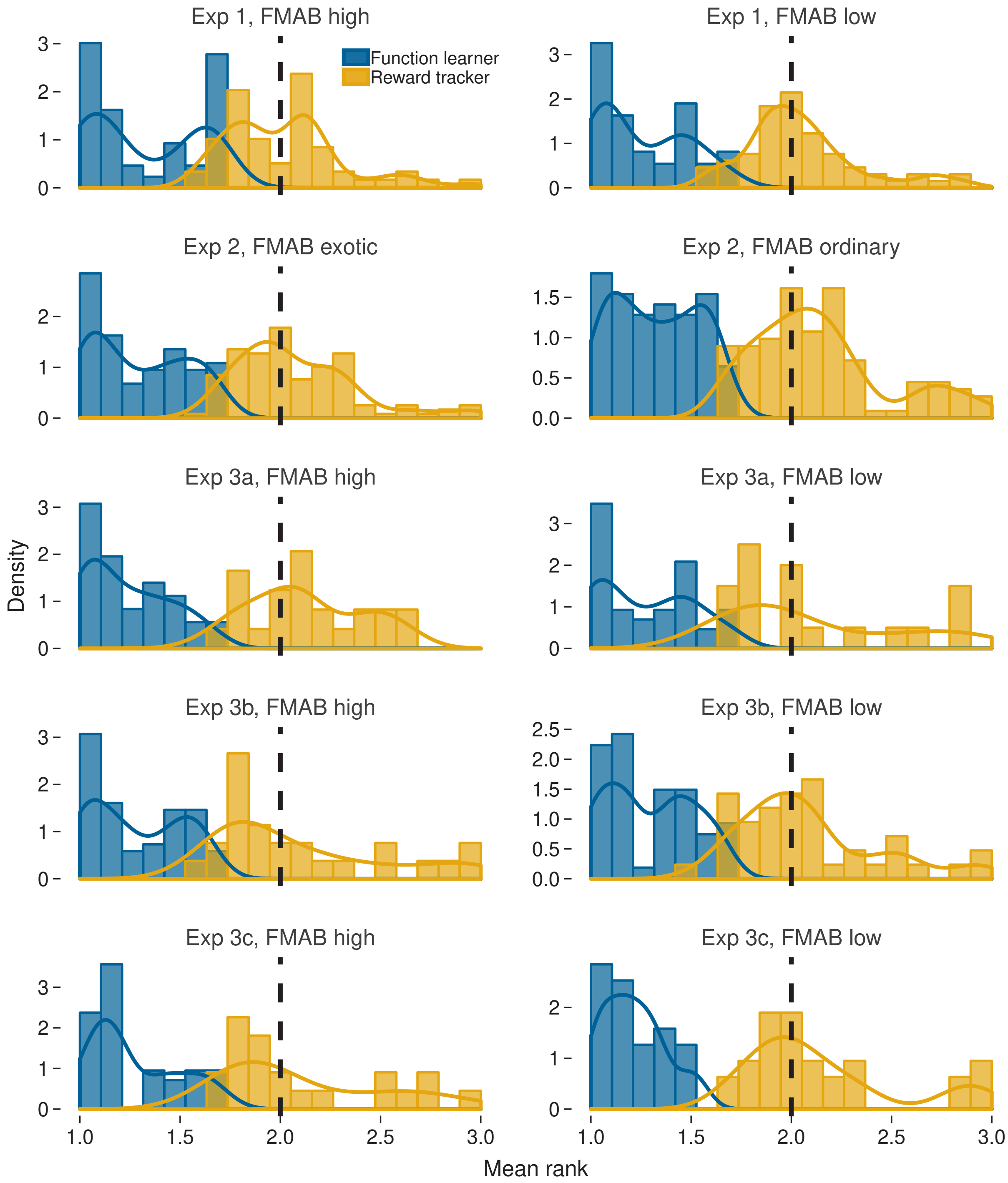


Figure B1. Histograms and densities of performance in functional knowledge task in Experiment 1 and Experiment 2. We classify each participant as a function learner if the mean rank of chosen options is significantly below random performance level of 2 (dashed vertical line), and as a reward tracker otherwise. Overall, performances of participants do not cluster around random performance, instead they span the whole range and the densities of the two groups show a good level of separation, suggesting that classification is meaningful.