

**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

Word count: 14380

Author Note

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.

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

### Abstract

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 two 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.

*Keywords:* Novelty; Decision making; Reinforcement learning; Function learning; Exploration-exploitation

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

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. 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. Monkeys can trade reward for novel information (Blanchard, Hayden, & Bromberg-Martin, 2015), and humans prefer novel stimuli to predictable ones (Berlyne, 1970) and newly-packaged goods over the same goods in old packaging (Steenkamp & Gielens, 2003). Novelty has also been found to attract attention in both children and adults (Nunnally & Lemond, 1974), and to bias the retrieval of episodes from memory such that higher value is attached to novel episodes (Carpenter & Schacter, 2016).

The tendency to seek out novel options can be beneficial: the reward a novel option provides 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; Whittle, 1980), and therefore, heuristic solutions are frequently employed. While not optimal, some heuristic strategies have been shown to work well.

One 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, thereby encouraging exploration of lesser-known options (e.g. Kakade & Dayan, 2002). This account resonates well with empirical findings suggesting 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, 2018; Speekenbrink & Konstantinidis, 2015; R. C. Wilson, Geana, White, Ludvig, & Cohen, 2014).

Contrary to the many findings suggesting that animals, including humans, 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: presenting an object repeatedly leads people to prefer this object over novel ones (Zajonc, 2001). Similarly, in self-directed learning people tend to choose options with known outcomes more frequently (Markant, Settles, & Gureckis, 2016).

To explain both novelty seeking and novelty avoidance, 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 induces novelty seeking. If a novel option is introduced in an environment where options are mostly not rewarding, this induces novelty avoidance. We expand upon this account of seemingly disparaging results. Specifically, our account is informed by the observation that options in real-world scenarios tend to come with features beyond their shared environment. For instance, 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 can be expected to be highly-rewarding as well, and thus should be approached. If the novel option is similar to non-rewarding options, it can be expected to be of poor quality, and thus should be avoided. 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. This requires agents to learn the function which relates features to rewards and people are indeed adept at learning functions. 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, 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 models of reinforcement learning as well as empirical evidence suggest that uncertainty may play a crucial role in how people choose among novel and time-honored options, over and above feature-based generalization. For example, 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. This changes the nature of exploration-exploitation dilemma subtly – exploration is 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, because 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, whilst keeping the uncertainty guided choices. Whether to approach or avoid a novel option depends on whether your functional knowledge predicts that it is good or bad, and whether approaching it helps improve your functional knowledge. In our previous work we have proposed a model that has these characteristics. The 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 this functional knowledge (Schulz, Konstantinidis, & Speekenbrink, 2018; Stojic, 2016; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018).<sup>1</sup> We have shown that people’s decisions are indeed guided by option-specific features expressed as abstract concepts (Analytis, Kothiyal, & Katsikopoulos, 2014; Stojic, 2016; Stojic, Analytis, & Speekenbrink, 2015; Wu, Schulz, Garvert, Meder, & Schuck, 2018), or embedded in the location of options (Wu, Schulz, Speekenbrink, et al., 2018), as well as shared features which relate to rewards in option-specific ways (Schulz, Konstantinidis, & Speekenbrink, 2018). We have also consistently found evidence for the use of sophisticated choice strategies that take into account 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 two pre-registered 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.

This work provides a bridge between human function learning and reinforcement learning, which have generally been studied in isolation. We believe that addressing both simultaneously is crucial for advancing knowledge about both topics. Function learning has

---

<sup>1</sup>See also Acuna and Schrater (2009) and Borji and Itti (2013) for related earlier work on human structure learning and decision making.



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, Bussemeyer, & 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 important insights into human experiential decision-making more generally. 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 when 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).

### **The feature-based multi-armed bandit task**

In the 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

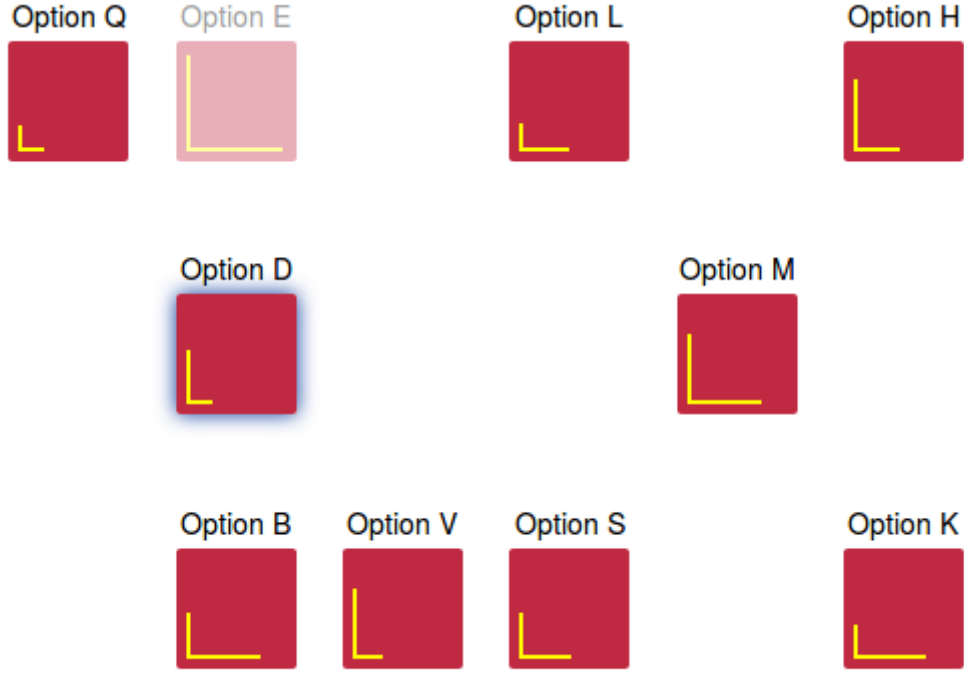
rewards associated to each option depend on the observable features through an initially unknown function which can be learned through experience. As in other multi-armed bandit tasks, learning requires participants to trade off between exploration – choosing options which reduce uncertainty about the function – and exploitation – 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 this novel option, we can discern whether participants exhibit functional generalization and functional uncertainty guidance when deciding to choose the novel option or not.

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 this general contextual multi-armed bandit. A related choice task with multiple dimensions has also been used to study the dynamics of attention in decision making tasks (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 use to tackle the task. 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



*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 we used is 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 10<sup>th</sup> option appears on the 41<sup>st</sup> trial in a randomly chosen position that was previously empty. In the illustration, the novel option is Option E 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 41<sup>st</sup> and 70<sup>th</sup> trial, participants have to estimate expected rewards of each option and express how confident they are in their estimates.

about the reward function. We contrast this to a reward tracking strategy which ignores feature information altogether and generalizes solely from experienced rewards.

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 differences between the strategies in terms of generalization also show in 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 is determined by 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. This 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 this prior to a posterior distribution over possible functions based on observed inputs (features) and outputs (rewards). Gaussian process regression assumes outputs  $y$

are generated from a function  $f$  over (multidimensional) inputs  $\mathbf{x}$  and additional noise  $\epsilon$ :

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

As a Bayesian technique, prior belief about the function  $f$  is formalized as a prior distribution over the possible functions. This prior distribution is defined as a Gaussian process:

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

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

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

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

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

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, this 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 this 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 an RBF kernel for a similarity-based account (Lucas et al., 2009), or as finding the appropriate combination of kernels (Schulz et al., 2017).

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

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

where the signal variance  $\sigma_f^2$  reflects the average distance of the function away from its mean and the length-scale  $\lambda$  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, closest 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). More importantly, 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 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 prediction’s uncertainty is crucial to guided exploration. The variance of this 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 this 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

$$u_{j,t} = m_{j,t} + \beta \sqrt{v_{j,t}}, \quad (4)$$

e.g., as the sum of the posterior predictive mean reward and a multiple ( $\beta$ ) of the uncertainty about the mean reward (the posterior predictive standard deviation). For a normally distributed variable, this 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 reward (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., 2018; R. C. Wilson et al., 2014). However, as we found most evidence for the UCB rule in 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:

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

Note that this 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. This 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 this 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

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

with a common variance  $\sigma_\epsilon^2$  but an option-specific mean  $\mu_j$ . The option-specific means are assumed to be drawn from a common higher-level Normal distribution

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

where  $\mu$  is the average reward over all options, and  $\tau^2$  the variance of the option-specific means. The model is completed with prior distributions for  $\mu$ ,  $\sigma$ , and  $\tau$ , 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  $\mu_j$ , the common mean

$\mu$ , and the variances  $\sigma^2$  and  $\tau^2$ . At any time  $t$ , this 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  $\mu$  (i.e., the expected reward for a novel option is thought to be the posterior mean of  $\mu$ ). 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. This strategy corresponds to trying out each option a few times and then deciding on the one with the highest utility. Because the hierarchical model also infers the distribution from which option’s average 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 this higher-level distribution. Importantly, as this 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,  $M_{\text{age}} = 37.1$  and  $SD_{\text{age}} = 10.5$ ) through Amazon’s Mechanical Turk (<http://mturk.com>) online labor market (Crump, McDonnell, & Gureckis, 2013; Paolacci & Chandler, 2014). There were 97 participants in both the FMAB high value and 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 was comprised of 70 trials in total. The same 9 options, each characterized by two features, were provided as

a choice set until the 41<sup>st</sup> 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 this option as the *novel option*.

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  $\epsilon_k^t$  was drawn from a Gaussian distribution with a mean of 0 and a variance of 4.

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 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 could therefore corroborate choice data evidence for the predictions. Hence, in

Option M



20
Confidence?

*Figure 2.* Participants completed an estimation task on the 41<sup>st</sup> and the 70<sup>th</sup> trial, after the choice but before receiving feedback on the reward. They had to estimate the expected reward of each option and express how confident they are in their estimates. This illustration shows the estimation task for the FMAB conditions. 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.

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 the confidence in their 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 assure 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. The equation 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.

**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). To distinguish those who do learn the function and those who do not, we



*Figure 3.* Illustration of a trial in the functional knowledge task. 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.

therefore developed a *functional knowledge task*, which acts as a model-free test whether participants have learned the reward function.

After the main bandit task, participants in the FMAB conditions only completed the functional knowledge task. The task consisted of 25 trials in which they had to choose between three new options without receiving feedback about the resulting reward (Figure 3). We designed the choice triplets such that there was always a best, a medium, and a worst option (for further details on construction of these items, see Appendix A). Participants who learned the function to some extent were expected to achieve better-than-chance performance in this task. We used participants’ mean performance in our preregistered classification procedure to distinguish between *function learners* and *reward trackers*. We classified participants who achieved better-than-chance performance as function learners, and participants who performed at chance-level or worse as reward trackers (for further details, see Appendix A).

**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 worker identification numbers using the Psiturk library.

Participants read the instructions after providing informed consent. We explained 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. 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 appear. We explained in detail how their earning would be computed 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 they were sent back to the instructions if they answered any of the questions incorrectly.<sup>2</sup>

Participants had a maximum of 60 seconds on every trial to select an option. To prevent participants from rapidly clicking through the experiment, they could not register their choice during the first second. Following each choice, reward feedback was displayed for two seconds, after which the task automatically continued to the next trial. 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 randomized between participants. The features (lines) were not displayed in the MAB conditions. We informed the FMAB participants that features were helpful, but did not inform them about the underlying function, which they had to infer 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

---

<sup>2</sup>There is moderate evidence that failing attention questions is correlated with performance in the FMAB conditions in both experiments as assessed 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  $H_0$ ), but no evidence of such an effect in the MAB conditions ( $BF_{10} = 0.27$ ).

task. The novel option appeared in one of the remaining 9 cells, selected at random, smoothly transitioning in transparency over a period of three seconds.

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 this task on the 41<sup>st</sup> and the 70<sup>th</sup> 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 (Figure 3). Immediately before the task, we instructed them that they would have to choose between new options on each trial and that they would not receive reward feedback on their choices, but that their final earnings would nevertheless 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 not labeled. 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 41<sup>st</sup> trial.<sup>3</sup>

**Analyses.** 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.

For hypotheses concerning participants’ choices of novel options in a single trial we used the contingency table Bayes factor of Jamil et al. (2017), with an independent

---

<sup>3</sup>We examined differences in choosing the novel option on trial 41 between those participants who indicated that they had noticed (29% in Experiment 1 and 27% in Experiment 2) and those who indicated they did not notice the novel option on trial 41. In Experiment 1 there was no evidence for the difference ( $BF_{10} = 0.12$ ), while there was a strong evidence in Experiment 2 ( $BF_{10} = 1.8 \times 10^4$ ). As we did not preregister hypotheses for this variable, and given that it might be hard to truthfully remember precisely when one noticed the appearance of the novel option, we did not analyze this variable in detail.



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 hierarchical estimation procedure (e.g. Kruschke, 2014). We used a non-centered probit parameterization and our priors of group-level means and standard deviations were informed by the model simulation results. 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 predictions are directional we used one-sided tests by truncating the prior above or below zero, with the null hypothesis being no difference and the alternative hypothesis a difference in the predicted direction. We used a symmetric, non-truncated prior whenever we had a non-directional hypothesis, and explicitly indicate when this was the case. More details about the statistical analyses can be found in Appendix A.

**Data and code availability.** All project files are publicly available at the Open Science Framework website: <https://osf.io/c8u9t/> (Stojic, Schulz, Analytis, & Speekenbrink, 2018b). This includes the behavioral data, the code used for our model simulations and data analysis. The preregistration document can be found at <https://osf.io/upj76/> (Stojic, Schulz, Analytis, & Speekenbrink, 2018a).

## Predictions

We generated *a priori* predictions by simulating the behavior of both the GP-UCB and the BH-UCB model in our task. We preregistered these predictions before data collection commenced (Stojic et al., 2018a).

To apply the GP-UCB model to our task, we determined the RBF hyperparameters

( $\sigma^2$  and  $\lambda^2$ ) 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., 2018a).

The simulation results confirm that the GP-UCB learns the reward function during the first 40 trials and is therefore able to 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). Importantly, this pattern is only predicted by the GP-UCB model, but not 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 41<sup>st</sup> trial, the GP-UCB model keeps selecting the high value option and its uncertainty reduces, while it ignores the low novel option for which uncertainty remains at a high level. This is 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).

Directly after its introduction on trial 41, the BH-UCB model chooses the high and low novel option with the same probability (Figure 4A, right panels). This is intuitive as the BH-UCB model cannot distinguish between novel options before they have been tried. It therefore 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

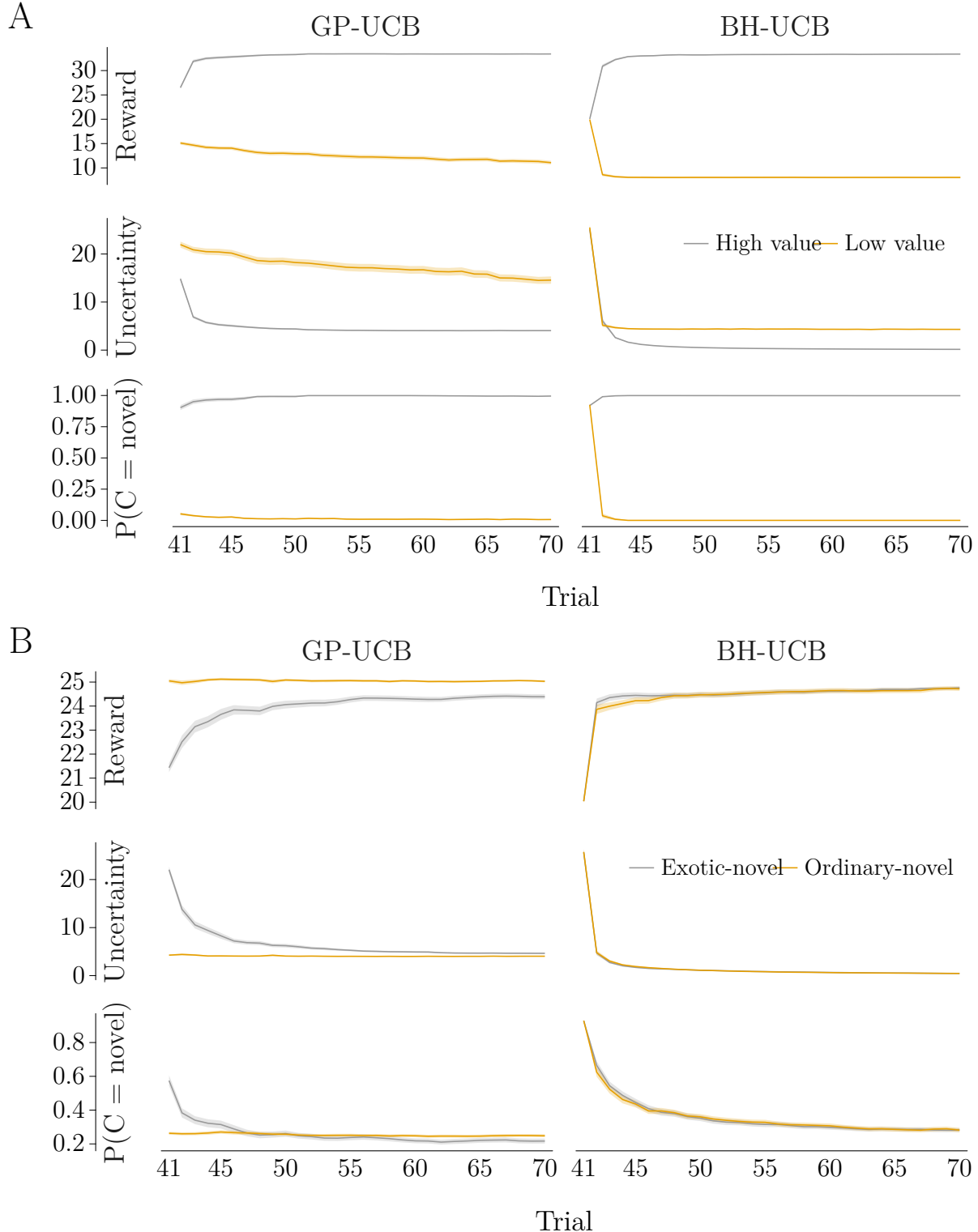


Figure 4. Simulation results for the GP-UCB and BH-UCB model for trials 41 to 70 in Experiment 1 (A) and in Experiment 2 (B). For each experiment we show the inferred reward (mean of the posterior predictive distribution, top row) and uncertainty for the novel option (variance of the posterior predictive 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 value aimed to capture a representative participant.

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 this happens depends on the magnitude of the exploration parameter, taking longer for smaller values of the exploration parameter (not shown in Figure 4A, see Stojic et al., 2018a).

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 41<sup>st</sup> trial (Stojic et al., 2018a). Contrary to this 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,  $BF_{10} = 0.07$ .<sup>4</sup>

Instead of emerging immediately on the 41<sup>st</sup> trial, the predicted difference arose from the 42<sup>nd</sup> trial onwards (Figure 5A). On the 42<sup>nd</sup> 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 novel condition eventually started choosing the novel option more often than participants in the low novel condition on trial 47 ( $BF_{10} = 10.93$ ) and onwards (Figure 5D). Importantly, as predicted, this difference in choice proportion between the high and low novel option emerged later than in the FMAB conditions.

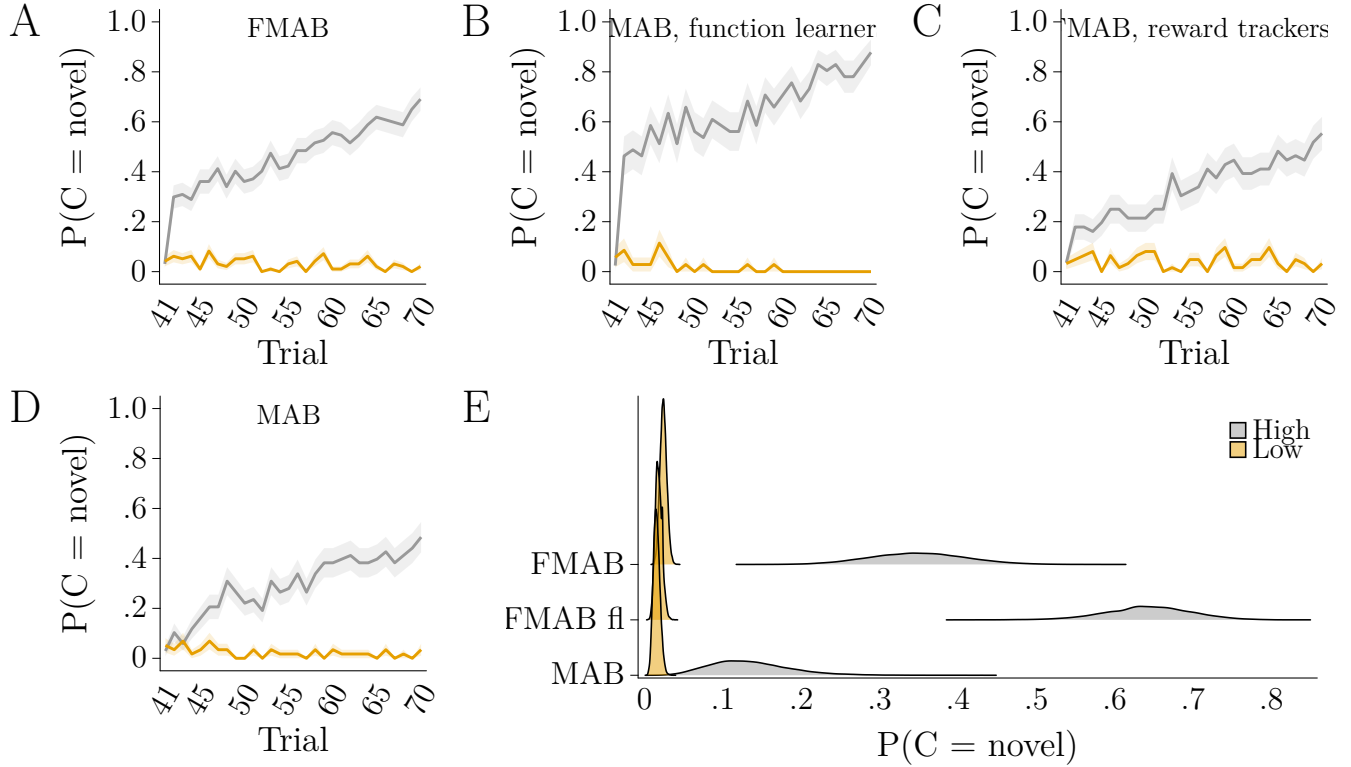
---

<sup>4</sup>Examining the choices only of participants who indicated at the end of the experiment that they had noticed the novel option on trial 41 we found the same result, i.e. no evidence for a difference in choice proportions between the two conditions ( $BF_{10} = 0.28$ ).

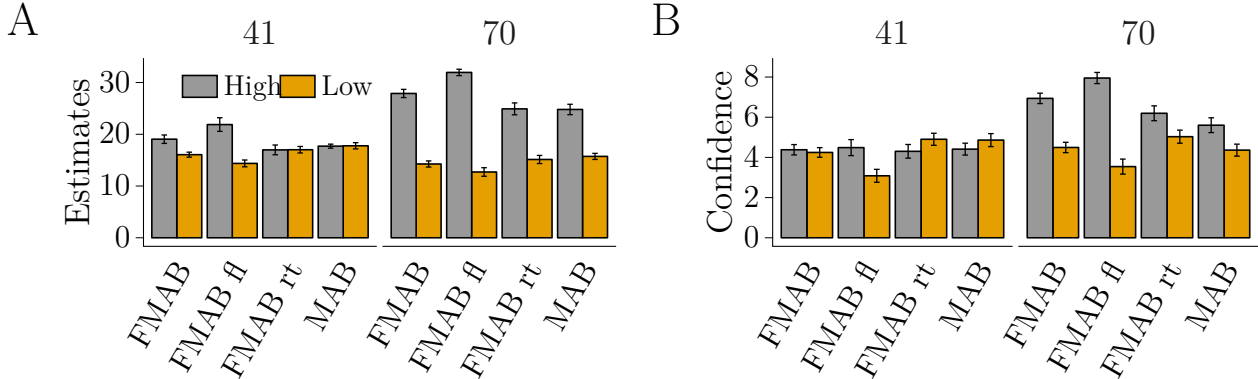
Moreover, 48.5% of participants did not choose the novel option at all in the FMAB low value condition, while only 21.6% of participants did so in the high value condition; a substantial difference with  $BF_{10} = 658$ .

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% in environments with linear reward functions (Stojic, 2016). As our predictions are geared towards function learners, and there are more reward trackers than function learners in the FMAB conditions, we also 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 predicted difference again arose from the 42<sup>nd</sup> trial onwards (Figure 5B), with a larger effect than for the FMAB conditions overall – in 42<sup>nd</sup> 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 picking up the high novel option, with strong evidence for a larger proportion of choices allocated to the high novel option (25%) than for the low novel option (0%) starting from trial 52 ( $BF_{10} = 2978$ ; Figure 5C), a pattern strikingly similar to that observed in the MAB conditions (Figure 5D).

As an alternative to examining single trials, we also used a Bayesian hierarchical model to estimate the probability of choosing the novel option on all 30 trials after its introduction (i.e., trial 41 to 70; see Figure 5E). The resulting posterior distributions of the probability that participants choose the high and low novel option showed a clear separation between the FMAB high and low novel 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



*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 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 high and low value option in similar proportions at the beginning, but as predicted, 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 FMAB conditions. In all four figures, lines reflect average choice proportions across participants, while filled bands reflect standard errors of the means. (E) Densities of posterior distributions over the probability of choosing the high or low novel option, estimated with a Bayesian hierarchical model for trials 41 to 70. Distributions show a clear difference for both FMAB conditions as a whole and function learning subgroup. In contrast, the difference between the MAB conditions is substantially smaller.



*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 novel 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 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 data from estimation task in those trials.

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 over the low novel 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.<sup>5</sup> According to the GP-UCB model predictions, FMAB participants – and function learners in particular – should estimate the high novel option to have higher reward than the low novel option on trial 41 (Figure 4A, top left

<sup>5</sup>The estimation task data passed our preregistered sanity checks. Our predictions that on trial 70 estimated values would be closer to options’ expected rewards and confidence ratings higher the more times the participants chose them was indeed true for both the FMAB ( $BF_{10} = 2.3 \times 10^9$  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  $H_0$ ).

panel). By contrast, following the BH-UCB model’s predictions, both MAB participants and reward trackers in the FMAB condition should evaluate both novel options as roughly equal (Figure 4A, top right panel). Indeed, on trial 41 FMAB participants correctly estimated the value of the high novel option (19.05) to be higher than the low novel (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 novel option (21.88) to be higher than the low novel 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 novel and 17.02 in the low novel condition,  $BF_{01} = 5.10$ , non-directional  $H_1$ ). By trial 70, as predicted by the BH-UCB model, MAB participants and reward trackers explored the novel options and exhibited the predicted differences in valuations of 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 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,  $BF_{10} = 8.6 \times 10^{26}$ ).

For FMAB participants and function learners in particular, the GP-UCB model predicts that confidence in estimated reward on trial 41 will be higher for the high novel option than the low novel option (Figure 4A, middle left panel). This prediction 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 novel option (4.49) than the low novel 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 rated their confidence equal (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 both models, confidence in the high novel option increased and was higher than for the low novel option (Figure 6B, trial 70 panel), more so for the FMAB participants (6.94 and 4.49,  $BF = 7.5 \times 10^7$ ) and function learners (7.95 and 3.54,  $BF = 8.6 \times 10^{11}$ ) than MAB participants (6.20 and 5.03,  $BF = 2.39$ ) and reward trackers (5.60 and 4.36,  $BF = 3.60$ ).

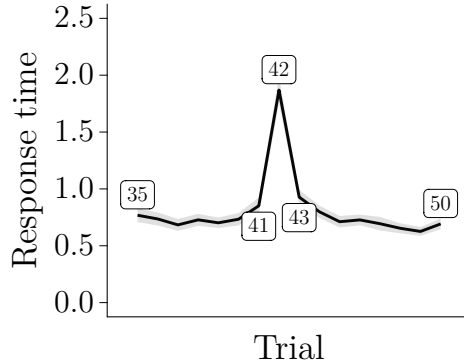
## Discussion

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 value of the high novel option was higher than that of the low novel option. More interestingly, 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 condition 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 ratings did not differ between the two options. Consequently, while they eventually discovered the high novel option, they did so much later than FMAB participants and particularly function learners.

A key preregistered hypothesis concerned the 41<sup>st</sup> trial, where we expected FMAB participants to choose the novel high value option more frequently than the novel low value option. This prediction was not confirmed by the results, but we found a strong effect on

the 42<sup>nd</sup> trial, as well as when comparing all trials from the 41<sup>st</sup> trial onwards. Moreover, data derived from the estimation task underpinned these results further. Why did participants only start considering the novel option after the 41<sup>st</sup> trial?

Despite our efforts to make the introduction of the novel option salient by smoothly increasing its opacity during a period of 3 seconds, participants might not have noticed its appearance on 41<sup>st</sup> trial. By the end of trial 40, many participants might have already settled on their next choice, registering this soon after the forced 1 second holdout. At that point, the novel option would still have been half-transparent, likely hindering its detection. Indeed, in the five trials preceding the 41<sup>st</sup> trial, the mean number of chosen options was 2.36 ( $SE = 0.10$ ) for the FMAB participants, of which 51.03% ( $SE = 3.59\%$ ) chose the same option on trial 40 as on trial 41. Their mean choice response time (excluding the one second holdout) during this period was 0.74s ( $SE = 0.04$ ) – too fast for the novel option to become fully visible. The introduction of the estimation task immediately after their choice on the 41<sup>st</sup> trial, which asked participants to make inferences for all options including the novel one, forced participant to pay attention to the novel option, allowing it to become part of the consideration set. The FMAB participants slowed down substantially on the 42<sup>nd</sup> trial, taking 1.87s ( $SE = 0.12$ ) to choose an option (Figure 7), and were much faster again during the five following trials (mean 0.79s,  $SE = 0.04$ ). Although a slowing of response time could be due to the preceding estimation task, this is unlikely to be the sole contributing factor: FMAB participants on average took 1.21s ( $SE = 0.01$ ) when they chose the novel option for the first time between trial 41 and 70, while they took 0.76s ( $SE = 0.0002$ ) for choosing the other options ( $BF_{10} = 2.7 \times 10^6$ , non-directional  $H_1$ ). For brevity we do not report results for the MAB conditions, where the results were qualitatively similar, but with slightly smaller differences. Taken together, these additional analyses suggest that the one-trial delay in the predicted differential preference between the high and low value novel options is likely due to participants not considering the novel option on trial 41; only after it was completely visible, and after the estimation task



*Figure 7.* Mean response times on trial 35 to trial 50. There is a large increase in response times on trial 42 in the FMAB conditions, where a large proportion of participants chose the novel option. One second holdout is not included in the response times and we removed outlier trials where participants took longer than three standard deviations above the mean response time over all choices in the dataset.

focused their attention to the novel option, did the novel option enter the consideration set.

To conclude, we found clear evidence for the predicted functional generalization effect. This shows 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. Confidence in predictions has received relatively 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, 2017; Folke, Jacobsen, Fleming, & De Martino, 2017). We found evidence that function learners track uncertainty about 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. This finding paves the way for Experiment 2, where we examined whether people are eager to approach more uncertain options.

## Experiment 2: Functional uncertainty guidance

The second preregistered experiment investigated functional uncertainty guidance, 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 experienced 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 experienced 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 produced the exact same reward in expectation. Crucially, uncertainty about the reward 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 options 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 sole 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 the predictions before collecting the data (Stojic et al., 2018a). The results of the simulation show that after introduction of the novel option on the 41<sup>st</sup> trial, the GP-UCB model chooses the exotic-novel option with a higher probability than the ordinary-novel option (Figure 4B, 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 4B, 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 view 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 have a strong theoretical commitment to the RBF kernel. Participants may not rely solely on an RBF kernel, but 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 pre-registered 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 function learners are not too few, whilst 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 both experiments simultaneously, a primary hypothesis for this experiment again focused on the 41<sup>st</sup> trial. As in Experiment 1, there was again no evidence for the predicted difference in choice proportions on this particular 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,  $BF_{10} = 0.25$ ).<sup>6</sup> In the MAB condition, 4.9% of the participants chose the novel option on trial 41.

Since the procedure was exactly the same as in Experiment 1, the same issues with participants’ attention to the novel option likely arose in this experiment as well, which was supported by additional analyses.<sup>7</sup> We therefore proceeded to explore participants’ behavior from the 42<sup>nd</sup> trial onwards. On the 42<sup>nd</sup> 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, this difference was expected to vanish rapidly. Behavioral data shows differences in choice proportions ranging from 4% to 12% over a longer period until trial 52, when the differences disappear. This extension of a relative preference over a prolonged period of time is likely due to participants’ slower learning as compared to the GP-UCB model

---

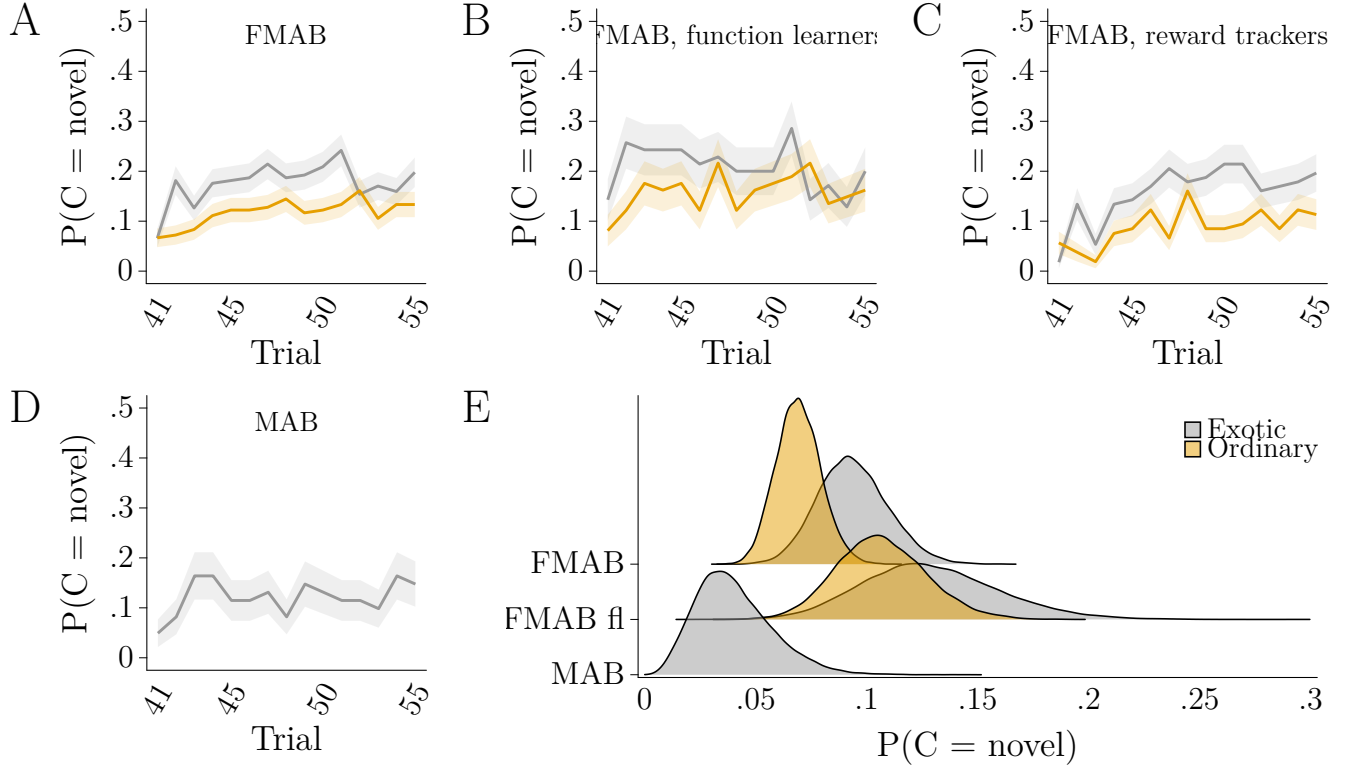
<sup>6</sup>Examining the choices only of participants who indicated at the end of the experiment that they had noticed the novel option on trial 41 we found the same result, no evidence for a difference in choice proportions between the conditions ( $BF_{10} = 0.22$ ).

<sup>7</sup>Reaction time characteristics were approximately the same as in Experiment 1 (Figure 7). FMAB participants took 1.84s ( $SE = 0.08$ ) to choose an option on the 42<sup>nd</sup> trial and they were faster during the five preceding trials, 0.81s ( $SE = 0.03$ ), and during the five following trials, 0.79 s ( $SE = 0.03$ ). FMAB participants on average took 1.21s ( $SE = 0.08$ ) when they chose the novel option for the first time on any trial between trial 41 to trial 70, while they took 0.73s ( $SE = 0.001$ ) for choosing the other options ( $BF_{10} = 2.7 \times 10^{17}$ , non-directional  $H_1$ ).

(Figure 8A). 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 of function learners as in Experiment 1. On the 42<sup>nd</sup> trial, function learners showed a similar differential preference for the exotic-novel option (25.7%) compared to the ordinary-novel option (12.2%) as FMAB participants generally, however resulting in a smaller Bayes factor of  $BF_{10} = 1.36$ , which is 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 used a Bayesian hierarchical model to estimate the probability of choosing the novel option across more than a single trial (Figure 8E). As we expected effects to diminish over trials, we focused this 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 8E). 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 probability estimates 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 exotic-novel condition was 12.7% (95% CI [7.4, 19.4]) and for the ordinary-novel 10.6% (95% CI [6.9, 14.9]), with 73% of probability estimates for exotic-novel condition larger than the ordinary-novel estimate and 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 probability estimates for the FMAB exotic-novel condition being larger than those for the MAB condition and a mean difference of 5.4%.





*Figure 8.* 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 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) In contrast, reward trackers in FMAB conditions choose both novel options in similar proportions at the beginning, 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 the FMAB exotic-novel and ordinary-novel conditions and MAB novel condition, estimated by a Bayesian hierarchical model for trials 41 to 55. They 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.

**Expected rewards and uncertainty.** Next, we examined participants' estimation task data to test whether participants' beliefs aligned with their choices (Figure 9A and 9B).<sup>8</sup> In line with the predictions from the GP-UCB model, on trial 41, FMAB participants estimated the ordinary-novel option to have a higher mean reward (18.54) than the exotic-novel option (15.75),  $BF_{10} = 24741$ . This difference was even larger when focusing solely on the function learners (19.49 for the exotic-novel and 15.37 for the ordinary-novel option,  $BF_{10} = 1637$ ). For reward trackers, there was no evidence that estimates of their expected rewards for 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 predicted by the GP-UCB model, 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 moderate evidence that their confidence ratings for the exotic-novel (4.59) were equal to the ratings for the ordinary-novel option (4.29),  $BF_{01} = 4.43$ .

Participants' confidence for the exotic-novel option increased substantially from trial 41 (3.84) to 70 (5.55),  $BF_{10} = 7.3 \times 10^6$ , indicating that once they tried out the exotic-novel option, their knowledge of its expected reward improved. As expected, 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 9A and 9B).

---

<sup>8</sup>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  $BF_{10} = 6.4 \times 10^{17}$ ).

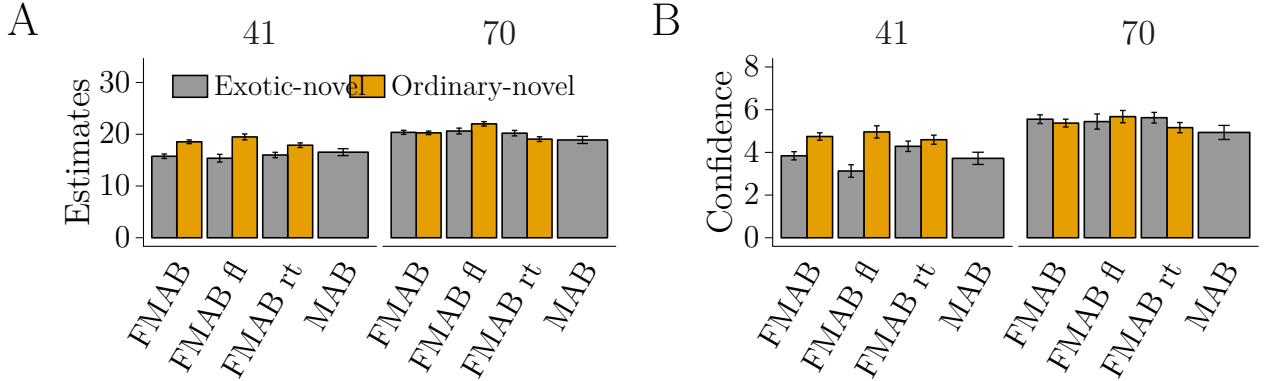


Figure 9. Estimated reward of the novel option and associated confidence in Experiment 2 for the estimation task on trial 41 and 70. (A) Estimated average reward of the novel option. FMAB participants and function learners in FMAB conditions (FMAB fl) underestimate the exotic-novel option on trial 41 but this 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, and this difference disappears again 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.

## 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 of the exotic-novel option on the 42<sup>nd</sup> trial, and a small overall preference in the period from the 41<sup>st</sup> to the 55<sup>th</sup> trial. As in Experiment 1, while our predictions did not hold on trial 41, we found similar evidence that participants were in the exploitation mode, responding quickly before the slowly appearing novel option was completely visible. 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.

Interestingly, our findings more closely followed the simulations with a Radial Basis Function kernel than would be expected if participants had extrapolated more linearly (e.g. Busmeyer et al., 1997). Using an RBF kernel, our model predicted a relatively small difference in choice proportions between the exotic-novel and ordinary-novel options, but

this 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 (cf. Hoffmann, von Helversen, & Rieskamp, 2016).

While our results indicate that participants integrated uncertainty into their decision process, it may have played a less prominent role, leading to moderate rather than strong differences in choice proportions. Results from the MAB condition, where participants rarely ever chose the novel option, indicate that this might be the case. Such a small probability of choosing the novel option can be accounted for by relatively small values of the  $\beta$  parameter in the UCB choice rule. However, it is less clear whether this holds for the FMAB conditions as well. Previous research has shown that participants can adapt the extent to which they rely on uncertainty in the environments they encounter (e.g. Behrens, Woolrich, Walton, & Rushworth, 2007). Since the benefit of incorporating uncertainty into decisions is greater in the FMAB conditions, it is likely that it played a more important role than in the MAB condition.

## General discussion

People constantly encounter a myriad of different options. This leads to a steady stream of decisions between novel and tried-and-tested options. Traditional models of reinforcement learning struggle to cope 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 information-rich and realistic settings,

identifying and choosing options that are not only novel but also good. This leads to the question of how this adaptive feature of human intelligence is accomplished.

We have put forward a model that combines functional generalization with uncertainty guidance to describe participants’ responses in the face of novelty to explain parts of this puzzle. This 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. Within two preregistered experiments, we used simulations of our model and a competing model which lacks the function learning component to generate qualitative predictions which we compared to participants’ behavior in a feature-based multi-armed bandit task. In the first experiment, we found that functional generalization can lead to both the seeking out of and the shunning away from novel options if their associated 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.

We found further support for functional generalization and uncertainty guidance by analyzing participants’ estimates of expected rewards and their uncertainty about all the options. 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  $\beta$  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 was that some of the predicted effects did not occur immediately on the 41<sup>st</sup> trial in which the novel option was first introduced. Rather, they occurred a trial later. We believe that this delay was due to participants not immediately noticing the novel option upon its introduction. As the novel option slowly became visible over a period of three seconds, whilst participants could register their choice after one second, many participants may have had made their choice before the novel option was fully visible. This interpretation is supported by additional analyses showing that participants did not slow down on the trial in which the novel option was introduced and often chose the same option as in the preceding trial. After the estimation task asked participants to consider all options, both old and novel, the expected effects did occur. The extent to which this can be attributed to the estimation task, or to the novel option becoming fully visible during the estimation task, is unclear. Regardless, it is likely that the novel option was not part of the considered set of options on trial 41.

Even if participants were fully aware of the novel option on the 41<sup>st</sup> trial and were somehow cautious in choosing it, we believe that this would not tarnish our account. Because participants could not have acquired direct information about the novel option without choosing it, there would have been relatively little change in the novelty of the option a trial later, where our predictions were confirmed. In the absence of further information about the novel option, theories which do not include functional generalization and uncertainty guidance will struggle to explain our results, whether they occurred on the 41<sup>st</sup> trial or shortly after it.

### **Individual differences and strategy selection**

As in prior research using the FMAB paradigm (Stojic, 2016), our results indicate that not everyone was willing or able to learn the feature-reward function; based on our

functional knowledge test, roughly 40% of participants were classified as function learners. Because the initial choice set was relatively small and participants were not explicitly told about the introduction of a novel item or the functional knowledge task, they may have opted for the less demanding reward-tracking strategy. Whilst this could reflect stable individual differences in ability or motivation, it could also reflect a cost-benefit arbitration between strategies (Kool, Gershman, & Cushman, 2017; Payne et al., 1993). Whether or not functional generalization and uncertainty guidance are worth the effort likely depends on environmental characteristics such as the size of the choice set, or the stationarity of rewards, for example. If the initial choice set is small and no new options are introduced, functional generalization may not be worth the cognitive effort. Ignoring the feature values whilst trying options may only give a small loss compared to inferring the feature-reward function, or even diminish loss if prior beliefs about the reward function are incorrect. If choices affect the reward function, planning becomes even more complex (Hotelling, Navarro, & Newell, 2018). In some domains, such as consumer goods, choice sets are likely to increase in average value over time, as manufacturers strive to deliver the best product for the lowest price. In those situations, knowledge of the feature-reward function in low-rewarding regions is unlikely to ever be useful. In other domains, where the quality of options diminishes with the number of times they are chosen (e.g., domains under the influence of resource depletion, see Pirolli & Card, 1999), the overall quality of a choice set may diminish over time. In those situations, it may be valuable to learn about the feature-reward function in low-rewarding regions, as suboptimal options within the current choice set may become the best available later on. If the choice between a function-learning or reward-tracking strategy is based on such prior expectations, it should be possible to steer this choice by manipulating people’s prior expectations. Alternatively, strategy selection can be viewed as a higher-level reinforcement learning problem itself (Lieder & Griffiths, 2017; Rieskamp & Otto, 2006; Stojic, Olsson, & Speekenbrink, 2016). For example, Lieder and Griffiths (2017) proposed a theory of strategy selection according to

which agents learn which strategy has the best cost-benefit trade-off. Focusing more explicitly on reinforcement learning strategies, Daw, Niv, and Dayan (2005) proposed that the arbitration between model-based and model-free reinforcement learning strategies is based on the relative uncertainty with which these strategies predict rewards. Identifying why some people employ a function-learning strategy, while others a reward-tracking strategy will be a valuable line of future research.

### **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 (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. 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 sample more experiences in consequential regions where feature values are likely to be rewarding. Accordingly, people are confident about their functional knowledge in that region of the space, whereas they remain less knowledgeable about low-rewarding feature values. Such relative ignorance is adaptive as long as choice



sets are not suddenly biased towards low-rewarding options.

In our study we purported that people are guided by both functional generalization and uncertainty. The idea that active function learning is geared towards uncertainty reduction is a compelling working hypothesis. Uncertainty reduction is a recurring theme in theories of intrinsically or extrinsically motivated information seeking (Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Kagan et al., 2002), however, little is known about how it plays out in active function learning tasks. In a recent study, Jones, Schulz, Meder, and Ruggeri (2018) showed that active function learners chose stimuli according to the same UCB rule as in the present study. They sampled items in regions with high output values, even though they were explicitly instructed to learn the entire function. These results provide converging evidence that even in the absence of extrinsic motivation, people sample using uncertainty. But they do so whilst focusing on regions where functional knowledge is expected to be more valuable, rather than on reducing experienced incongruities (Hebb, 2005) or on shedding light on information-gaps (Loewenstein, 1994). Because of this focus, a local (i.e. exemplar) representation of the function where accuracy is perceived to matter most may be more useful than a more global (e.g. linear) representation of the function. As such, the bias towards consequential regions may in fact modulate how the function is learned in the first place.

### **Limitations, future directions and applications**

People’s exploration is affected by the task horizon. As predicted by rational models, they decrease the amount of exploration with the number of choices left (R. C. Wilson et al., 2014). Currently, the GP-UCB model does not incorporate such a dynamic exploration policy. It would be straightforward to include it in a heuristic manner, by making the exploration parameter dependent on 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 uncertainties (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, as proposed by Guez et al. (2013), has firmer normative grounds than simply decreasing an exploration parameter over time. Notably, Krusche, Schulz, Guez, and Speekenbrink (2018) found empirical evidence for human planning in a challenging foraging task.

Positioning novelty within our functional generalization and uncertainty guidance framework may provide new insights into the very concept of novelty. 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 “truly novel”. We believe 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 shelf 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 for different domains. Transfer of learning is currently at the frontier of machine learning research (e.g. Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016; Wang et al., 2016) and would constitute an exciting extension of our framework.

Our results regarding functional generalization and uncertainty guidance may prove

useful in areas such as marketing, where clear predictions about when a customer will explore a novel product are important (Hirschman, 1980; Riefer, Prior, Blair, Pavey, & Love, 2017). What should newly-introduced products look like in order to maximize their market share? Market analysts often rely on conjoint analysis and other preference elicitation techniques to settle on attractive product features (Green & Srinivasan, 1990; Netzer et al., 2008). Similar to multi-attribute perspectives in choice (Keeney & Raiffa, 1993), these methods assume that people have well-formed and stable preferences, ignoring that people may be actively learning their preferences. Our theory suggests that uncertainty may be a crucial factor in enticing someone to try new products. In repeated choice contexts, people will try products that are not only predicted to be good, but which also appeal to their curiosity. The extent of uncertainty guidance may differ between people and choice contexts, and this may potentially explain why some consumers are early adopters and others are followers (Mahajan, Muller, & Srivastava, 1990; Rogers, 2010).

### **Concluding remarks**

In summary, we believe that our model constitutes as a powerful and expressive account of human behavior in the face of novelty. Beyond novelty, integration of function learning and decision making allows revisiting familiar problems from a new perspective and opens up new avenues of research, as highlighted in our discussion. 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.

## 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). Multi-attribute utility models as cognitive search engines. *Judgment and Decision Making*, *9*, 403–419.
- 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.
- 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. (2017). Confidence modulates exploration and exploitation in value-based learning. *bioRxiv*. doi: 10.1101/236026
- 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). Curran Associates, Inc.
- Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*, 1–27.
- Bunzeck, N., & Düzel, E. (2006). Absolute Coding of Stimulus Novelty in the Human Substantia Nigra/MTA. *Neuron*, *51*, 369–379. doi: 10.1016/j.neuron.2006.06.021

- Busmeyer, 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.
- 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.
- 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., Busmeyer, 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.
- 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.
- 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.
- Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, 17, 585–593.
- Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: new developments with implications for research and practice. *The Journal of Marketing*, 3–19.
- 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.
- Hebb, D. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press.
- 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
- Hotaling, J. M., Navarro, D. J., & Newell, B. R. (2018). Skilled bandits: Learning to choose in a reactive world.
- 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.
- Jones, A., Schulz, E., Meder, B., & Ruggeri, A. (2018). Active function learning. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 580–585). Austin, TX.
- 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.
- Kagan, J., et al. (2002). *Surprise, uncertainty, and mental structures*. Harvard University Press.
- 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. *Journal of the American Statistical Association*, 90, 773–795.
- Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge University Press.
- 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.
- 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.
- 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.
- Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*, 75.
- 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.
- 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.
- Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., . . . others (2008). Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters*, *19*, 337–354.
- 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.
- Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*, 643–675.
- R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Reichel, C. M., & Bevins, R. A. (2008). Competition between the conditioned rewarding effects of cocaine and novelty. *Behavioral Neuroscience*, *122*, 140–150.
- 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
- 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 Research*, 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
- 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., 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. (2018). Searching for rewards like a child

- means less generalization and more directed exploration. *bioRxiv*. doi: 10.1101/327593
- 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(5), 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., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2018a). *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. (2018b). *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/>
- 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*.
- 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.

## Appendix

### Functional knowledge task

Participants in the FMAB conditions completed an additional functional knowledge task. 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 greater weight. Denoting dominating, middle and dominated option as  $x$ ,  $y$  and  $z$ , the feature values were generated as follows:

- For the easy type items, we first drew the feature values for the dominating option. Once these feature values were known, the intervals for sampling feature values of the middle option were generated. We constructed the feature values of the dominated 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 difficult triplets in a similar manner, but the intervals for generating the middle and dominated option were closer to the dominating option, making the middle and dominated option more similar to the dominating 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.  $x_2 \sim U(y_2 - .05, \min(.7, y_2 + \frac{w_1}{w_2}(y_1 - x_1)))$
4.  $z_1 \sim U(\max(x_1, y_1) + .05, .8)$
5.  $z_2 \sim U(\max(x_2, y_2) + .05, .8)$

- The “weight comparison” items consisted of a dominating option with a small value



of the feature with the largest weight and a large value of the other feature. The middle option had exactly the opposite pattern, thereby creating a diagnostic pair for detecting whether people have learned which feature is more predictive. The dominated option had two large feature values.

$$1. x_1 \sim U(0.25, 0.35), x_2 \sim U(0.7, 0.8)$$

$$2. y_1 \sim U(0.7, 0.8), y_2 \sim U(0.25, 0.35)$$

$$3. z_1 \sim U(0.7, 0.8), z_2 \sim U(0.7, 0.8)$$

## 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 two experiments: participants’ choices and response times in the bandit task and functional knowledge task, their estimates and confidence ratings in the estimation task, as well as age, gender, and whether they had noticed the appearance of the novel option.

**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 41<sup>st</sup> 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 favour 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.

**Statistical tests.** We use Bayes factors to quantify the relative evidence the data provides in favour 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 hierarchical estimation procedure (e.g. Kruschke, 2014) 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 FMAB conditions in Experiment 1 and from trial 41 to 55 in Experiment 2. Hierarchical model treats participants as members of a group, taking into account group probability distribution when estimating individual probabilities. This leads to more realistic posterior distribution of 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 and here our choice was informed by 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,  $p = 0.01$ ) for low value and  $\mu \sim N(2.21, 1)$  ( $p = 0.99$ )

for high value condition. For the MAB conditions the same priors were  $\mu \sim N(-1.83, 1)$  ( $p = 0.03$ ) for low value and  $\mu \sim N(2.75, 1)$  ( $p = 0.99$ ) for high value condition. In Experiment 2 for the FMAB conditions (and function learner subgroups) priors were based on simulation results for trials 41 to 46 where models show a difference,  $\mu \sim N(-0.38, 1)$  ( $p = 0.35$ ) for exotic-novel and  $\mu \sim N(-0.61, 1)$  ( $p = 0.27$ ) for ordinary-novel condition. For the MAB conditions the priors were  $\mu \sim N(0.20, 1)$  ( $p = 0.58$ ) for exotic-novel and  $\mu \sim N(0.12, 1)$  ( $p = 0.55$ ) for 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 of each chain, after discarding the first 40000 of burn-in samples of each chain. We confirmed that all chains successfully converged by visually inspecting the chains 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 of 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 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 performance in the functional knowledge task. We compared the mean rank of participants’ choices (rank 1 being the best and 3 being the worst alternative) across all 25 choices in the task to the mean rank of a person choosing randomly, which equals 2. The null hypothesis was that there is no difference, while the alternative hypothesis was that the mean rank is lower than two. If there was strong evidence ( $BF_{0A} > 10$ ) that a participant’s mean rank is below 2, we classified the participant as a function learner, and as a reward tracker otherwise.

Few sanity-check hypotheses involved testing for the relationship between the number of times an option was chosen and 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 has 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 regarding dependence of choice performance and number of failed attention checks.