bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Putting bandits into context: How function learning supports decision making

Eric Schulz
University College London

Emmanouil Konstantinidis
University of New South Wales

Maarten Speekenbrink
University College London

We introduce the contextual multi-armed bandit task as a framework to investigate learning and decision making in uncertain environments. In this novel paradigm, participants repeatedly choose between multiple options in order to maximise their rewards. The options are described by a number of contextual features which are predictive of the rewards through initially unknown functions. From their experience with choosing options and observing the consequences of their decisions, participants can learn about the functional relation between contexts and rewards and improve their decision strategy over time. In three experiments, we find that participants' behaviour is surprisingly adaptive to the learning environment. We model participants' behaviour by context-blind (mean-tracking, Kalman filter) and contextual (Gaussian process regression parametrized with different kernels) learning approaches combined with different choice strategies. While participants generally learn about the context-reward functions, they tend to rely on a local learning strategy which generalizes previous experience only to highly similar instances. In a relatively simple task with binary features, they mostly combine this local learning with an "expected improvement" decision strategy which focuses on alternatives that are expected to improve the most upon a current favourite option. In a task with continuous features that are linearly related to the rewards, they combine local learning with a "upper confidence bound" decision strategy that more explicitly balances exploration and exploitation. Finally, in a difficult learning environment where the relation between features and rewards is non-linear, most participants learn locally as before, whereas others regress to more context-blind strategies.

Keywords: Function Learning; Decision Making; Gaussian Process; Multi-Armed Bandits; Reinforcement Learning

Introduction

Imagine you recently arrived in a new town and need to decide where to dine tonight. You have visited a few restaurants in this town before and while you have a current favourite, you are convinced there must be a better restaurant out there. Should you revisit your current favourite again tonight, or go to a new one which might be better, but might also be worse? This is an example of the exploration-exploitation dilemma (e.g., Cohen, McClure, & Yu, 2007; Laureiro-Martínez, Brusoni, & Zollo, 2010; Mehlhorn et al., 2015): in order to benefit the most, should you exploit your current but incomplete knowledge to pick an option you think is best, or should you explore something new and improve upon your knowledge in order to make better decisions in the future? While exploration is risky, in this case it is not blind. Over the years, you have visited many restaurants and you know for instance that better restaurants generally have more customers, a good ambiance, and are not overly cheap. So you walk around town, noting of each restaurant you pass how busy it is, how nice it looks, the price of the items on the menu, etc. At the end of a long walk, you finally sit down in a restaurant; one you never visited before but predicted to be best based on numerous features such as neighbourhood, clientéle, price, and so forth.

The exploration-exploitation dilemma tends to be studied with so-called multi-armed bandit tasks, such as the Iowa

Eric Schulz and Maarten Speekenbrink, Department of Experimental Psychology, University College London, London, UK; Emmanouil Kontantinidis, School of Psychology, University of New South Wales, Sydney, Australia.

This research is supported (ES) by the UK Centre for Training in Financial Computing and Analytics.

Correspondence concerning this article should be addressed to Eric Schulz, Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, UK. E-mail: e.schulz@cs.ucl.ac.uk.

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

gambling task (e.g., Bechara, Damasio, Tranel, & Damasio, 2005; Steyvers, Lee, & Wagenmakers, 2009). These are tasks in which people are faced with a number of options, each having an associated average reward. Initially, these average rewards are unknown and people can only learn about the reward of an option by choosing it. Through experience, people can learn which are the good options and attempt to accumulate as much reward as possible over time. However, as our restaurant example above shows, many real-life situations are richer than such simple multi-armed bandit tasks. Options tend to have numerous features (e.g., number of customers and menu prices in the restaurant example) which are predictive of their associated reward. With the addition of informative features, the decision problem can be termed a *contextual* multi-armed bandit (henceforth CMAB; Li, Chu, Langford, & Schapire, 2010). While these kinds of tasks are ubiquitous in daily life, they are rarely studied within the psychological literature. This is unfortunate, as CMAB tasks encompass two important areas of cognition: experience-based decision making (Barron & Erev, 2003; Hertwig & Erev, 2009) and function learning (DeLosh, Busemeyer, & McDaniel, 1997; Kalish, Lewandowsky, & Kruschke, 2004). Both topics have been studied extensively (see e.g., Newell, Lagnado, & Shanks, 2015, for an overview), but in isolation.

Learning and decision making within contextual multi-armed bandit tasks generally requires two things: learning a function that maps the observed features of options to their expected rewards and a decision strategy that uses these expectations to choose between the options. Function learning in CMAB tasks is important because it allows one to generalize previous experiences to novel situations. For example, it allows one to predict the quality of a new restaurant from experiences with other restaurants with a similar number of customers and a similarly priced menu. The decision strategy is important because not only should you attempt to choose options that are currently most rewarding, but you should also take into account how much you can learn in order to make good choices in the future. In other words, you should take into account the exploration-exploitation trade-off, where exploration here means learning about the function that relates features to rewards.

In what follows, we will describe the contextual multi-armed bandit paradigm in more detail and propose several models to describe how people may solve these tasks. We will then assess how participants perform within three different variants of a CMAB task. We will show that participants are able to learn within the CMAB and are best-described by sensitive exploration-exploitation behaviour that adapts to the situation at hand and by locally approximating the true underlying function (Lucas, Griffiths, Williams, & Kalish, 2015; Srinivas, Krause, Kakade, & Seeger, 2009). In summary, we make the following 3 contributions:

1. We introduce the contextual multi-armed bandit as a psychological paradigm combining both function learning and decision making.
2. We model learning behaviour by parametrizing Gaussian Processes with different kernel functions. Gaussian Processes are a powerful tool for regression problems and generalize important psychological models previously proposed.
3. We show that participants sensibly choose between options according to their expectations while locally learning about the underlying functions.

Contextual multi-armed bandits

A contextual multi-armed bandit task can be formalized as a game in which on each round, an agent is presented with a context (a set of features) and a set of options which each offer an unknown reward. The context can contain general features that apply to all options (e.g., the country the restaurants are in) or specific features that apply to single options (e.g., the exact menu and its price). The agent's task is to choose those arms that will accumulate the highest reward over all rounds of the game. The rewards are stochastic, such that even if the agent had complete knowledge of the task, a choice would still involve a kind of gamble. In this respect, choosing an option can be seen as choosing a slot machine (a one-armed bandit) to play, or choosing an arm of a multi-armed bandit. After choosing an option in a round, the agent receives the reward of the chosen option but is not informed of the rewards that could have been obtained from the other options. The expected rewards associated to each option depend on the context through an unknown function. For an agent who ignores the context, the task would appear as a restless bandit task (e.g., Speekenbrink & Konstantinidis, 2015), as the rewards associated with an arm will vary over time due to the changing context. However, learning the function that maps the context to (expected) rewards will make these changes in rewards predictable and thereby choosing the optimal arm easier. In order to choose wisely, the agent should learn about the underlying function. Sometimes, this may require her to choose an arm which is not expected to give the highest reward on a particular round, but one that might provide useful information about the function, thus choosing to explore rather than to exploit.

Contextual multi-armed bandit tasks provide us with a scenario in which a participant has to learn a function and potentially maximize expected outputs of that function over time by making wise choices. They are a natural extension of both the classic multi-armed bandit task, which is a CMAB with an invariant context throughout, and the restless bandit task, which is a CMAB with time as the only contextual feature. While the CMAB is novel in the psychological literature, where few tasks explicitly combine function learning and experience-based decision making, there are certain

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

similarities with tasks used in previous research. For example, recent studies in experience-based decision-making provided participants with descriptions about the underlying distributions that generate rewards (e.g., Lejarraga & Gonzalez, 2011; Weiss-Cohen, Konstantinidis, Speekenbrink, & Harvey, 2016). Just as in the CMAB, this presents a naturalistic decision environment in which different sources of information (e.g., descriptions and participants' own experience) need to be integrated in order to choose between alternatives or courses of action.

Another related paradigm is multiple cue probability learning (MCPL, Kruschke & Johansen, 1999; Speekenbrink & Shanks, 2008) in which participants are shown an array of cues that are probabilistically related to an outcome and have to learn the underlying function mapping the cues' features to expected outcomes. Especially when the outcome is a categorical variable, such as in the well-known "Weather Prediction Task" (Gluck, Shohamy, & Myers, 2002; Speekenbrink, Channon, & Shanks, 2008) where participants predict the state of the weather ("rainy" or "fine") based on a set of "tarot cards", making a prediction is structurally similar to a decision between multiple arms (possible predictions) that are rewarded (correct prediction) or not (incorrect prediction). Just as in the CMAB, multiple-cue probability learning and probabilistic category learning tasks require people to learn a function which maps multiple cues or features to expected outcomes. An important difference however is that in these latter tasks there is a strong dependency between the arms: there is only one correct prediction, and hence there is a perfect (negative) correlation between the rewards for the arms. Whether a current choice was rewarded or not thus provides information about whether the non-chosen arms would have been rewarded. This dependency weakens the need for exploration, especially when the outcome is binary, in which case there is no need for exploration at all. In CMAB tasks, there is more need for exploration as the rewards associated to arms are, conditional on the context, generally independent. Knowing that a particular arm was rewarded does not provide immediate information whether another arm was rewarded. Another major difference is that MCPL tasks generally require participants to learn the whole function. In CMAB tasks, learning the function is only necessary insofar as it helps to make better decisions. To solve the exploration-exploitation dilemma, it may suffice to learn the function well only in those regions that promise to produce high rewards. Moreover, as we will see later, each arm can be governed by its own function relating context to rewards. To our knowledge, simultaneous learning of multiple functions has not previously been investigated.

Another area of related research comes from the associative learning literature, where it has been shown that context can act as an additional cue to maximize reward (cf Bouton & King, 1983; Gershman, Blei, & Niv, 2010). In one example of this, Gershman and Niv (2015) showed how the generalization of context (the average reward of options in an environment) can explain how participants react to novel options in the same environment, where a context of high rewards leads to a positive response to novel options, and a context with low reward rates to low responses to novel options. The CMAB paradigm introduced here is similar to such tasks but adds the additional feature of having to learn an underlying contextual function.

Models of learning and decision making

Formally, the CMAB is a game in which on each round $t = 1, \ldots, T$, an agent observes a context $s_t \in \mathcal{S}$ from the set $\mathcal{S}$ of possible contexts and has to choose an arm $a_t \in \mathcal{A}$ from the set $\mathcal{A}$ of all arms of the multi-armed bandit. Afterwards, the agent receives a reward

$$y_t = f(s_t, a_t) + \epsilon_t$$

and it is her task to choose those arms that will produce the highest accumulated reward $R = \sum_{t=1}^{T} y_t$ over all rounds. The function f is initially unknown and can only be inferred from the reward received after choosing an arm in a particular context.

To perform well in a CMAB task, an agent needs to learn a model of the function f from experience and use this model to predict the outcomes of available actions and choose the arm with the best outcome. We can thus distinguish between a learning component, formalized as a learning model which takes previous observations to estimate the outcomes of a function, and a decision or acquisition component that uses the learned model to determine the best subsequent decisions. These work together as shown in Algorithm 1 (see also Brochu, Cora, & De Freitas, 2010).

Algorithm 1 General CMAB-algorithm. A learning model $\mathcal{M}$ tries to learn the underlying function f by mapping the current expectations and their attached uncertainties to choices via an acquisition function acq.

Require: A model $\mathcal{M}$ of the function f, an acquisition function $\text{acq}_{\mathcal{M}}$, previous observations $\mathcal{D}_0 = \{\emptyset\}$
for $t = 1, 2, \ldots, T$ **do**
  Choose arm $a_t = \arg\max_{a \in \mathcal{A}} \text{acq}_{\mathcal{M}}(a | s_t, \mathcal{D}_{t-1})$
  Observe reward $y_t = f(s_t, a_t) + \epsilon_t$
  Update Augment the data $\mathcal{D}_t = (a_t, s_t, \mathcal{D}_{t-1})$ and update the model $\mathcal{M}$ **end for**

This formalization of an agent's behaviour requires us to capture two things: (a) a representation or model $\mathcal{M}$ of the assumed underlying function that maps the given context to expected outcomes and (b) an acquisition function $\text{acq}_{\mathcal{M}}$ that evaluates the utility of choosing each arm based on those expected outcomes and their attached uncertainties. Here,

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the model defines the learning process and the acquisition function the way in which outputs of the learned model are mapped onto choices. In the following, we will describe a number of instantiations of these two components.

Models of learning

Technically, a function is a mapping from a set of input values to a set of output values, such that for each input value, there is a single output value (also called a many-to-one mapping as different inputs can provide the same output). Psychological research on how people learn such mappings has generally followed a paradigm in which participants are presented with input values and asked to predict the corresponding output value. After their prediction, participants are presented with the true output value, which is often corrupted by additional noise. Through this outcome feedback, people are thought to adjust their internal representation of the underlying function. In psychological theories of function learning, these internal representations are traditionally thought to be either *rule-based* or *similarity-based*. Rule-based theories (e.g., Carroll, 1963; Koh & Meyer, 1991) conjecture that people learn a function by assuming it belongs to an explicit parametric family, for example linear, polynomial or power-law functions. Outcome feedback allows them to infer the parameters of the function (e.g., the intercept and slope of a linear function). This approach attributes a rich set of representations (parametric families) to learning agents, but tends to ignore how people choose from this set (how they determine which parametric family to use). Similarity-based theories (e.g., Busemeyer, Byun, Delosh, & McDaniel, 1997) conjecture that people learn a function by associating observed input values to their corresponding output values. When faced with a novel input value, they form a prediction by relying on the output values associated to input values that are similar to the novel input value. While this approach is domain general and does not require people to assume a parametric family a priori, similarity-based theories have trouble explaining how people readily generalize their knowledge to novel inputs that are highly dissimilar to those previously encountered.

Research has indicated that neither approach alone is sufficient to explain human function learning. Both approaches fail to account adequately for the finding that some functional forms, such as linear ones, are much easier to learn than others, such as sinusoidal ones (McDaniel & Busemeyer, 2005). This points towards an initial bias towards linear functions, which can be overcome through sufficient experience. They also fail to adequately predict how people extrapolate their knowledge to novel inputs (DeLosh et al., 1997). In order to overcome some of the aforementioned problems, hybrid versions of the two approaches have been put forward (McDaniel & Busemeyer, 2005). One such hybrid is the *extrapolation-association model* (EXAM, DeLosh et al., 1997), which assumes a similarity-based representation for interpolation, but extrapolates using simple linear rules. Although *EXAM* effectively captures the human bias towards linearity and predicts human extrapolations over a variety of relationships, it fails to account for the human capacity to generate non-linear extrapolations (Bott & Heit, 2004). The *population of linear experts model* (POLE, Kalish et al., 2004) is set apart by its ability to capture knowledge partitioning effects; based on acquired knowledge, different functions can be learned for different parts of the input space. Beyond that, it demonstrates a similar ordering of error rates to those of human learners across different tasks (McDaniel, Dimperio, Griego, & Busemeyer, 2009). Recently, Lucas et al. (2015) proposed Gaussian process regression as a rational approach towards human function learning. Gaussian process regression is a Bayesian non-parametric approach which unifies both rule-based and similarity-based theories of function learning. Instead of assuming one particular functional form, Gaussian process regression is based on a model with a potentially infinite number of parameters, but parsimoniously selects parameters through Bayesian inference. As shown by Lucas et al., it can explain many of the previous empirical findings on function learning. Following this approach, we will conceptualize function learning in a CMAB as Gaussian process regression. We contrast this with context-blind learning which tries to directly associate an option to an expected reward without taking the contextual features into account.

Contextual learning through Gaussian process regression. A Gaussian process (GP) is a stochastic process such that the marginal distribution of any finite collection of observations drawn from it is a multivariate Gaussian (see Rasmussen, 2006). Gaussian process regression is a non-parametric Bayesian approach (Gershman & Blei, 2012) towards regression problems and can be seen as a "rational" way to learn functions that adapts its own complexity to the data encountered (see Griffiths, Lucas, Williams, & Kalish, 2009). In the following, we will assume that the agents learns a separate function $f_j(s)$ that maps contexts s to rewards y for each arm j. A $\mathcal{GP}$ defines a distribution $p(f_j)$ over such functions, parametrized by a mean function $m_j(s)$ and a covariance function, also called kernel, $k_j(s, s')$:

$$m_j(s) = \mathbb{E}\left[f_j(s)\right] \tag{1}$$

$$k_j(s, s') = \mathbb{E}\left[(f_j(s) - m_j(s))(f_j(s') - m_j(s'))\right] \tag{2}$$

In the following, we will focus on the computations for a single option and suppress the subscripts j. Suppose we have collected rewards $\mathbf{y}_t = [y_1, y_2, \ldots, y_t]^\top$ for arm j in contexts $\mathbf{s}_t = \{s_1, \ldots, s_t\}$, and we assume

$$y_t = f(s_t) + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Given a $\mathcal{GP}$ prior on the functions

$$f(s) \sim \mathcal{GP}\left(m(s), k(s, s')\right). \tag{4}$$

the posterior over f is also a $\mathcal{GP}$ with

$$m_t(s) = \mathbf{k}_t(s)^\top(\mathbf{K}_t + \sigma^2\mathbf{I})\mathbf{y}_t \tag{5}$$

$$k_t(s, s') = k(s, s') - \mathbf{k}_t(s)^\top(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}\mathbf{k}_t(s') \tag{6}$$

where $\mathbf{k}_t(s) = [k(s_1, s), \ldots, k(s_t, s)]^\top$ and $\mathbf{K}_t$ is the positive definite kernel matrix $[k(s, s')]_{s,s' \in \mathcal{D}_t}$. The posterior variance can also be computed as

$$v_t(s) = k_t(s, s). \tag{7}$$

This posterior distribution can also be used to derive predictions about each arm's rewards given the current context, that are also assumed to be normally distributed.

A key aspect of a GP model is the covariance, or kernel function k. The choice of a kernel function corresponds to assumptions about the shape of the true underlying function. Among other aspects, the kernel determines the smoothness, periodicity, and linearity of the expected functions (c.f. Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2016). Additionally, the choice of the kernel also determines the speed at which a GP model can learn over time (Schulz, Tenenbaum, Reshef, Speekenbrink, & Gershman, 2015). The kernel defines a similarity space over all possible contexts. As such, a GP can be seen as a similarity-based (or exemplar) model of function learning. However, by first mapping the contexts s via the kernel into a "feature space", it is possible to rewrite the posterior mean of a GP as a linear combination of transformed feature values. From a psychological perspective, a GP model can in this way also be thought of as encoding "rules" mapping inputs to outputs.

A GP can thus be expressed as both an exemplar (similarity-based) model and a feature (rule-based) model, thereby unifying the two dominant classes of function learning theories in cognitive science (as described in Lucas et al., 2015).

Different kernels correspond to different psychological assumptions about how people approach function learning. We will compare 4 different kernels that together span a reasonable range of these assumptions. By choosing a *linear kernel*, the model corresponds directly to Bayesian linear regression. This kernel thus instantiates a relatively simple rule-based way of learning the underlying function, assuming it has a particular parametric shape, namely a linear combination of the contextual features. On the other side of the spectrum, an *Ornstein-Uhlenbeck process kernel* assumes very rough, unsmooth functions, so that the structure of the underlying function is effectively learned only very locally to the observations made. As this kernel hardly generalizes encountered features to other contexts, this kernel can be seen as instantiating a form of exemplar-based learning combined with sparse extrapolation. The *Mátern kernel* is parameterized to represent an average level of smoothness of the underlying function. Both the Mátern and Ornstein-Uhlenbeck kernels can be seen as more general function approximators than the linear kernel, but they differ in how local their inferences are. In general, kernels with higher correlations between distant points expect smoother functions, which means observing one point provides more information about other points. The *radial basis function kernel* (sometimes also called squared exponential or Gaussian kernel) postulates infinitely smooth functions and is probably the most frequently used kernel within the Gaussian process literature. In our setting, each kernel comes with 2 free parameters, a general scaling parameter θ_1 and a length-scale parameter θ_2 which determines how quickly dependencies diminish as points are further removed. The mathematical details of the 4 kernels, as well as an illustration of the way in which the 4 different kernels learn (i.e. update their prior distribution over functions to a posterior distribution) are provided in Table 1.

Context-blind learning. To assess the extent to which people take the context into account, we contrast the contextual learning models above with three context-blind learning models that ignore the features and focus on the average reward (and possibly its attached uncertainty) of each option over all contexts.

The *random model* picks each option with the same probability and constitutes as a simple baseline against which the other models can be compared.

The *Bayesian mean-tracking* model assumes that the average reward associated to each option is constant over time and simply computes a posterior distribution for the mean μ_j of each option j. Here, we will implement a relatively simple version of such a model which assumes rewards are normally distributed with a known variance but unknown mean and the prior distribution for that mean is again a normal distribution. This implies that the posterior distribution for each mean is also a normal distribution:

$$p(\mu_j|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t})$$

The posterior distribution can be computed through a mean-stable version of the Kalman Filter, which we will describe next. Here, the mean $m_{j,t}$ represents the currently expected outcome for a particular arm j and the variance $v_{j,t}$ represents the uncertainty attached to that expectation.

Unlike the Bayesian mean tracking model, which computes the posterior distribution of a time-invariant mean μ_j after each new observation, the Kalman filter is a suitable model for tracking a time-varying mean $\mu_{j,t}$ which we here assume varies over time according to a simple random walk

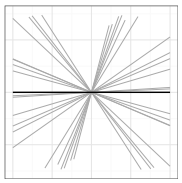
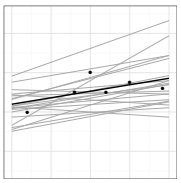
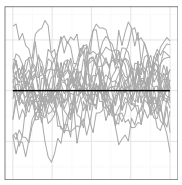
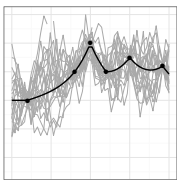
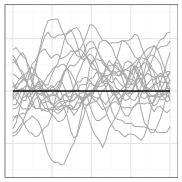
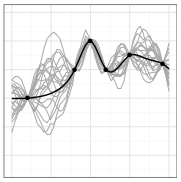
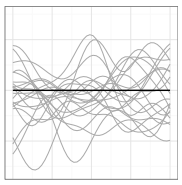
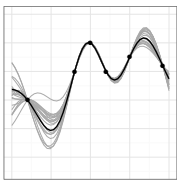
$$\mu_{j,t+1} = \mu_{j,t} + \zeta_t \qquad \zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2)$$

Such a Kalman filter model has been used to successfully describe participants choices in a restless bandit task (Speekenbrink & Konstantinidis, 2015) and has also been proposed

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1

Details of the different kernels used to model participants' learning. Mathematical details of each kernel are provided in the column labeled "kernel". Prior samples from each resulting Gaussian process for a one-dimensional input are shown in the "prior" column. The "posterior" column shows posterior samples of the functions from each Gaussian process after the same set of 6 observations (dots).

Kernel	Prior	Posterior
Linear $\theta_1(s-\theta_2)(s'-\theta_2)$		
Ornstein-Uhlenbeck $\theta_1 \exp\left(-\frac{\lvert s-s'\rvert}{\theta_2}\right)$		
Mátern 3/2 $\theta_1\left(1+\frac{\sqrt{3}\lvert s-s'\rvert}{\theta_2}\right)\exp\left(-\frac{\sqrt{3}\lvert s-s'\rvert}{\theta_2}\right)$		
Radial Basis $\theta_1 \exp\left(-\frac{(s-s')^2}{2\theta_2^2}\right)$		

as a model unifying many findings within the literature of context-free associative learning (Gershman, 2015). The posterior distribution of the mean is again a normal distribution

$$p(\mu_{j,t}|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t})$$

with mean

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t}[y_t - m_{j,t-1}] \tag{8}$$

where y_t is the received reward on trial t and $\delta_{j,t} = 1$ if arm j was chosen on trial t, and 0 otherwise. The "Kalman gain" term is computed as

$$G_{j,t} = \frac{v_{j,t-1} + \sigma_\zeta^2}{v_{j,t-1} + \sigma_\zeta^2 + \sigma_\epsilon^2}$$

where $v_{k,t}$, is the variance of the posterior distribution of the mean $\mu_{j,t}$ is computed as

$$v_{j,t} = [1 - \delta_{j,t} G_{j,t}][v_{j,t-1} + \sigma_\zeta^2] \tag{9}$$

Prior means and variances were initialized to $m_{j,0} = 0$ and $v_{j,0} = 1000$, while the innovation variance σ_ζ^2 and error variance σ_ϵ^2 were free parameters. The Bayesian mean tracking model is obtained from the Kalman filter model by setting the innovation variance to $\sigma_\zeta^2 = 0$, implying the underlying mean is not assumed to change over time.

Decision strategies

As the aforementioned models generate an expectation given the current context represented by a predictive distribution, we need a decision strategy defining how to choose among different arms given different estimates of their current predictive means and variances. In the psychological literature, popular decision rules that map current expectations onto choices are Luce's choice rule (Luce, 1963) and the ϵ-greedy rule (Sutton & Barto, 1998). These are simple rules that are only based on a single expectation for each option. In Luce's choice rule, the probability of choosing an option is roughly proportional to the current expectations, while the

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ϵ-greedy rule chooses the maximum-expectancy option with probability $1 - \epsilon$ and otherwise chooses with equal probability between the remaining options. These rules ignore the uncertainty about the formed expectations, while rationally, uncertainty should guide exploration. Here, we follow Speekenbrink and Konstantinidis (2015) and define a rich set of decision rules that explicitly model how participants trade off between expectations and uncertainty.

We will consider 4 different strategies to make decisions in a CMAB task based on the expected outcomes derived from the above learning models. The mathematical details of these are given in Table 2. The *upper confidence bound* (UCB) algorithm estimates a trade-off between the current expected value and the variance per option and chooses the option with the highest upper confidence bound; it has been shown to perform well in many real world tasks (Krause & Ong, 2011). The UCB rule has a free parameter c which determines the width of confidence interval (e.g., setting $c = 1.96$ results in a 95% credible set). The UCB-algorithm can be described as a selection strategy with an exploration bonus, where the bonus dynamically depends on the confidence interval of the estimated mean reward at each time point. It is sometimes also referred to as optimistic sampling as it can be interpreted to inflate expectations with respect to the upper confidence bounds (Srinivas et al., 2009). We will approximate the confidence interval for each option by a 95% credible set based on a normal distribution (so fixing $c = 1.96$).

Another decision strategy is the *probability of improvement* which calculates the probability for each arm to lead to an outcome higher than the best observed outcome so far (Kushner, 1964). Intuitively, this algorithm estimates the probability of one option to generate a higher utility than another option and has recently been used in experiments involving multi-attribute choices (Gershman, Malmaud, Tenenbaum, & Gershman, 2016).

The *expected improvement* is similar to the probability of improvement, but calculates the expected increase of outcomes for each arm compared to the maximum output seen so far (Mockus, Tiesis, & Zilinskas, 1978).

The fourth decision strategy we consider is the *probability of maximum utility* rule (Speekenbrink & Konstantinidis, 2015). This strategy chooses each arm according to the probability that it results in the highest reward out of all arms in a particular context. It can be seen as a form of probability matching (Neimark & Shuford, 1959) and can be implemented by sampling from each arm's predictive distribution once, and then choosing the arm with the highest sampled pay-off. Even though this acquisition function seems relatively simplistic at first, it can describe human choices in restless bandit tasks well (Speekenbrink & Konstantinidis, 2015). It is also closely related to Thompson sampling (May, Korda, Lee, & Leslie, 2012), which samples from the posterior distribution of the mean rather than the predictive distribution of rewards. Thus, while Thompson sampling "probability matches" the expected rewards of each arm, the probability of maximum utility rule matches to actual rewards that might be obtained.

All of these decision rules (apart from the Probability of Maximum Utility rule) are essentially deterministic. As participants' decisions are expected to be more noisy reflections of the decision rule, a softmax-transformation was used to transform the utilities into probabilities of choice:

$$p(a_t = j) = \frac{\exp\{\gamma \cdot \mathrm{acq}(a = j|s_t, \mathcal{D}_{t-1})\}}{\sum_{i=1}^{n} \exp\{\gamma \cdot \mathrm{acq}(a = i|s_t, \mathcal{D}_{t-1})\}} \tag{10}$$

The temperature parameter $\gamma > 0$ governs how consistent participants choose according to the values generated by the different kernel-acquisition function combinations. All free parameters were estimated by numerically maximising the likelihood of participants' decisions with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.

General CMAB task

In our implementation of the CMAB task, participants are told that they have to mine for "Emeralds" on different planets. Moreover, it is explained that –at each time of mining– the galaxy is described by 3 different environmental factors, "Mercury", "Krypton", and "Nobelium", that have different effects on different planets. Participants are then told that they have to maximize their production of Emeralds over time by learning how the different environmental factors influence the planets and choosing the planet they think will produce the highest outcome in light of the available factors. Participants were explicitly told that different planets can react differently to specific environmental factors. A screenshot of the CMAB can be seen in Figure 1.

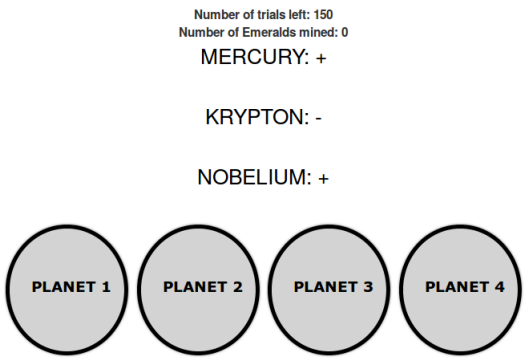


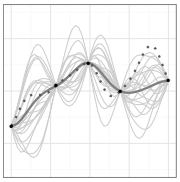
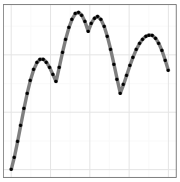
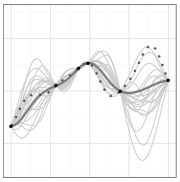
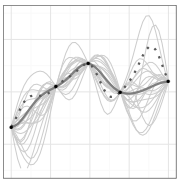
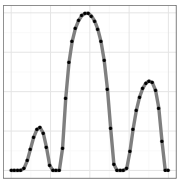
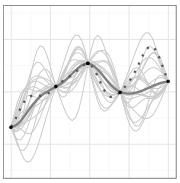
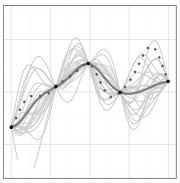
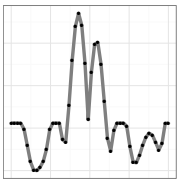
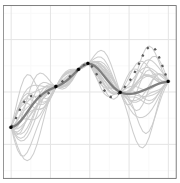
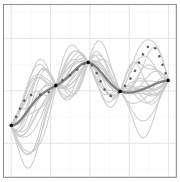
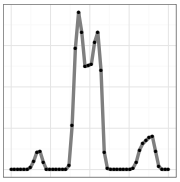
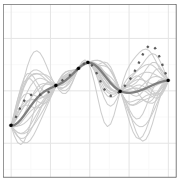
Figure 1. Screenshot of experiment.

As each planet responds differently to the contexts, they can be seen as arms of a bandit that are related to the context by different functions. The reward an option j provides is

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2

Different acquisition functions used to model participants' choices. Mathematical details are provided in the column "Acquisition function". Here, $m_{*,t}(s)$ denotes the posterior mean of the function for context s and action $*$ which is the action currently believed to be optimal. Examples are provided for a problem where each action corresponds to choosing a one-dimensional input, after which the associated output can be observed. Prior samples from a Radial Basis kernel are shown in the "Prior (time t)" column. The utility of each potential action according to each acquisition function is shown in the "Acquisition function" column. After choosing the action with the highest utility and observing the corresponding output, the Gaussian process is updated and used as a prior at the next time. Samples from this posterior are shown in the final column.

Acquisition function (acq($a = i \mid s_t, \mathcal{D}_{t-1}$))	**Prior (time t)**	**acq(x)$_k$**	**Prior (time $t+1$)**
Upper Confidence Bound: $m_{j,t}(s_t) + c\sqrt{v_{j,t}(s_t)}$			
Probability of Improvement $\Phi\left(\frac{m_{j,t}(s) - m_{*,t}(s)}{\sqrt{v_{j,t}(s)}}\right)$			
Expected Improvement $\left(m_{j,t}(s) - m_{*,t}(s)\right)\Phi(z) + \sqrt{v_{j,t}(s)}\phi(z)$, $z = \frac{m_{j,t}(s) - m_{*,t}(s)}{\sqrt{v_{j,t}(s)}}$			
Probability of maximum utility $P(f_j(s) + \epsilon_{j,t} > f_i(s) + \epsilon_{i,t}, \forall i \neq j)$			

given as

$$y_{j,t} = f(a_t = j, s_t) = f_j(s_t) + \epsilon_{j,t}$$

with $\epsilon_{j,t} \sim \mathcal{N}(0, 5)$. The task consists of 150 trials in which a random context is drawn and participants choose a planet to mine on. Which planet corresponded to which reward function f_j was determined at random before the start of the experiment.

The three experiments we present differed in the functions f_j and whether the environmental factors were binary or continuous. This is specified in more detail when describing the experiments. Source code for all experiments is available online.[1]

Experiment 1 : CMAB with binary cues

The goal of the first experiment was to test whether participants can learn to make good decisions in a CMAB task. For this purpose, we set up a relatively simple contextual bandit scenario in which the contexts consist of binary features.

Participants

Forty-seven participants (26 male) with an average age of 31.9 years ($SD = 8.2$) were recruited via Amazon Mechanical Turk and received \$0.3 plus a performance-dependent bonus of up to \$0.5 as a reward.

[1] https://github.com/ericschulz/contextualbandits

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Task

There were four different arms that could be played. In addition, three discrete variables, $s_{i,t}$, $i = 1, 2, 3$, were introduced as the general context. The three variables defining the contexts could either be on ($s_{i,t} = 1$) or off ($s_{i,t} = -1$). The outcomes of the four arms were dependent on the context as follows:

$$f_1(s_t) = 50 + 15 \times s_{1,t} - 15 \times s_{2,t}$$
$$f_2(s_t) = 50 + 15 \times s_{2,t} - 15 \times s_{3,t}$$
$$f_3(s_t) = 50 + 15 \times s_{3,t} - 15 \times s_{1,t}$$
$$f_4(s_t) = 50$$

On each trial, the probability that a contextual feature was on or off was set to $p(s_{i,t} = 1) = p(s_{i,t} = -1) = 0.5$. The functions f_j were deliberately designed such that the expected reward of each arm over all possible contexts is $\mathbb{E}[y_{j,t}] = 50$. This means that the only way to gain higher rewards than the average of 50 is by learning how the contextual features influence the rewards. More formally, this means that no arm achieves first-order stochastic dominance. Moreover, by including the context-independent fourth arm that returns the mean with added noise helps us to distinguish even further between learning and not learning the context: this arm has the same expected value as all the other arms but a lower variance and therefore second-order dominates the other arms. As such, a context-blind learner would be expected to prefer this arm over time.

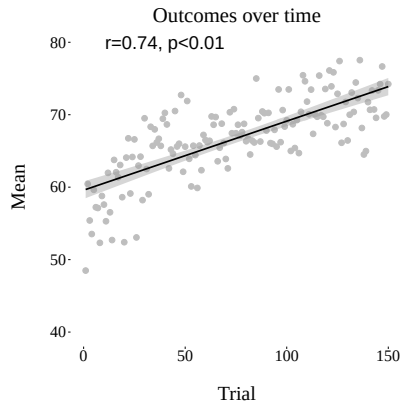
Procedure

As described above, participants were told that they had to mine for "Emeralds" on different planets. Moreover, it was explained that at each time each of the 3 different environmental factors could either be on (+) or off (-) and had different effects on different planets. Participants were told that they had to maximize the overall production of Emeralds over time by learning how the different elements influence the planets and then picking the planet they thought would produce the highest outcome, given the status (on or off) of the elements. It was explicitly noted that different planets can react differently to different elements. There were a total number of 150 trials.

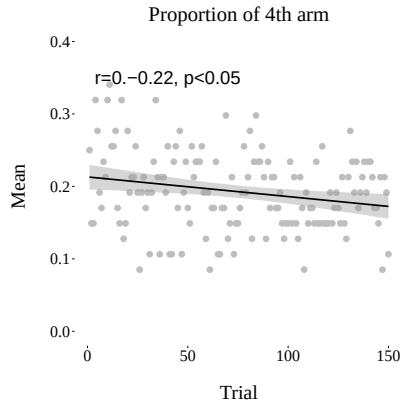
Results

Behavioral results. Participants learned to take the context into account and gained 66.78 points (SD=13.02) on average throughout the task. Average scores of participants were significantly above the chance level of 50, as confirmed by a one-sample t-test, $t(46) = 7.17$, $p < 0.01$.

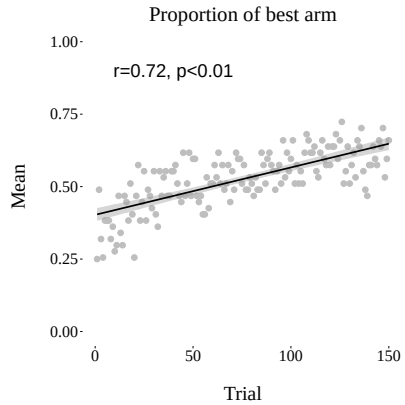
Over time, participants made increasingly better choices (see Figure 2a), as indicated by a significant correlation between the average score (over participants) and trial number, $r = 0.74$, $p < 0.01$. The proportion of participants choosing the non-contextual option (the option that did not respond to any of the contextual features, indicated as the 4th arm) decreased over time ($r = -0.22$, $p < 0.05$, Figure 2b), another indicator that participants learned the underlying functions. Finally, the proportion of participants choosing the



(a) Average scores per round.



(b) Proportion of 4th arm.



(c) Proportion of best arm.

Figure 2. Average overall score (a), mean score per round (c), and proportion of best arm chosen per round (c) for the binary CMAB task.

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

best option for the current context increased during the task ($r = 0.72$, $p < 0.01$, see Figure 2a).

Modelling results. To determine which combination of learning model and an acquisition function best captures participants' choices, we focus on three indicators of model fit. For each participant and model, we computed Akaike's "An Information Criterion" (AIC, Akaike, 1974), which penalizes model fit by model complexity. Lower AIC values indicate better model performance. For each model, we also counted the number of participants best-described by the model according to the AIC. Finally, we computed the "protected exceedance probability" (Rigoux, Stephan, Friston, & Daunizeau, 2014) for each model. This measure assumes that each model occurs with some frequency in the population of all participants and defines the protected exceedance probability as the probability that a particular model is more frequent than all the other models, whilst accounting for the possibility that there are truly no differences between the models in their frequency of occurrence.[2] The results are shown in Figure 3.

Overall, the best performing model was a GP learning model with the Ornstein-Uhlenbeck kernel paired with expected improvement as decision rule. This combination had the lowest average AIC (297), a protected probability that it is the most frequent model of $p = 0.24$, and described $n = 6$ participants best overall. Other models which described participants' behaviour well incorporated the same decision strategy, but used a Mátern 3/2 (mean AIC 298, $n = 5$, $p = 0.18$) or Radial Basis kernel (mean AIC 298, $n = 6$, $p = 0.16$). We can assess the performance of the learning models and acquisition functions separately by computing the marginal protected probability of exceedance. Marginalizing over over all acquisition functions, the Mátern 3/2 kernel was most likely to be the most frequent learning model ($p = 0.32$), followed by the Ornstein-Uhlenbeck kernel ($p = 0.26$). Marginalizing over all learning models, showed that the expected improvement is by far most likely to be the most frequent decision rule ($p = 0.61$).

The Ornstein-Uhlenbeck, and to a somewhat lesser extent the Mátern 3/2 kernel, assume relatively unsmooth functions. As a result, there seems to be only relatively sparse generalization from observations in one context to similar other contexts. The results of our cognitive modelling exercise thus imply that people learn rather locally and may have relied on a special form of exemplar learning combined with sparse extrapolation, effectively memorizing for each context and arm the observed rewards and averaging these to come to an estimate of the expected reward. By using an Expected Improvement decision strategy, they would compare the arm with the highest averaged rewards in a particular context to relatively unknown arms in that context, determining how probable these are to provide a higher reward and the magnitude of this improvement. This is in agreement with prior findings in more simple multi-attribute choice tasks (for example, Carroll & De Soete, 1991).

Experiment 2: Continuous-Linear CMAB

Experiment 1 indicated that people may have relied on a very local, exemplar-like learning strategy. As that experiment contained only 8 unique contexts, a memorization strategy is feasible. The goal of the second experiment was to assess whether such local learning extends to a situation with more unique contexts. In Experiment 2, we used the same task but with continuous rather than discrete features to comprise the contexts and a linear relation between contexts and rewards.

Participants

Fifty-nine participants (30 male) with a mean age of 32.4 (SD=7.8) were recruited via Amazon Mechanical Turk and received \$0.3 as a basic reward and a performance-dependent bonus of up to \$0.5.

Task and Procedure

The task was identical to that of Experiment 1, only this time the context contained continuous features with an underlying linear function mapping inputs to outputs:

$$f_1(s_t) = 50 + 3 \times s_{1,t} - 3 \times s_{2,t}$$
$$f_2(s_t) = 50 + 3 \times s_{2,t} - 3 \times s_{3,t}$$
$$f_3(s_t) = 50 + 3 \times s_{3,t} - 3 \times s_{1,t}$$
$$f_4(s_t) = 50$$

The values of the context variables $s_{j,t}$ were described numerically and sampled randomly from a uniform distribution $s_{j,t} \sim \mathcal{U}(-10, 10)$. Again, the expected value (over all contexts) for each planet was 50, so there was no first-order stochastically dominating arm, while the fourth arm achieved second-order stochastic dominance as the variance of its rewards was the lowest.

Results

Behavioral results. As in Experiment 1, participants were able to take the context into account. The average obtained reward across participants was 59.84 (SD = 9.41), significantly higher than chance, $t(58) = 7.17$, $p < 0.01$. However, two participants scored significantly less than expected even at chance level assessed by a simple t-test and were therefore excluded from the following analysis.

[2] To estimate this probability, we first estimated the log-evidence of each model by a Laplace approximation (based on the determinant of the Hessian matrix computed in the BFGS-optimization), and then used these as the likelihood of each model in the hierarchical Bayesian model of Stephan, Penny, Daunizeau, Moran, and Friston (2009); (see also Gershman, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

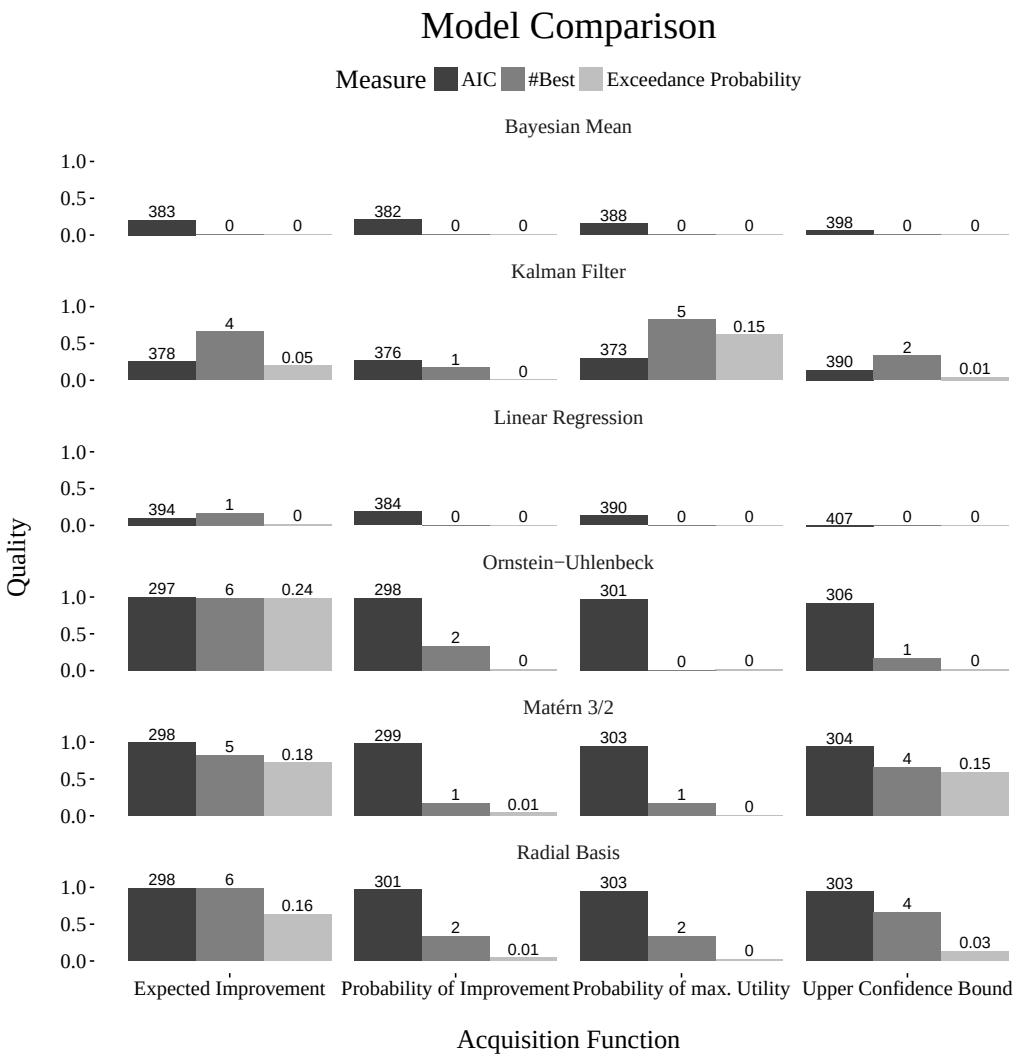


Figure 3. Results of model comparison for CMAB with binary cues when accounted for random model (mean AIC=415.89, 0 participants explained best, exceedance probability=0). Results were standardized to fit on one scale.

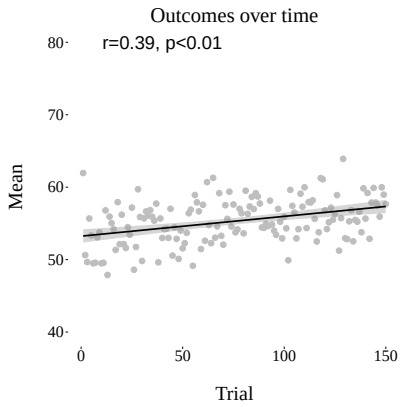
Performance increased over trials, $r = 0.39$, $p < 0.01$, although this was not as pronounced as in Experiment 1 (see Figure 4a).

The proportion of participants choosing the fourth arm did not decrease over time ($r = 0.05$, $p > 0.05$). Notice however that in this scenario with higher input variance (caused by more diverse input-output pairs), it can happen that the fourth option turns out to be the best option overall. The proportion of choosing the best option given the context significantly increased over trials ($r = 0.33$, $p < 0.01$, see Figure 4c).

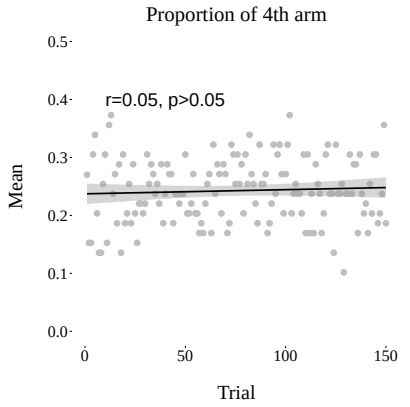
Modelling results. Modelling results are shown in Figure 5. The best model of participants' behaviour overall combined a GP learning strategy with an Ornstein-Uhlenbeck kernel with the Upper Confidence Bound decision strategy (mean AIC = 371, number of best fitting participants $n = 15$, protected exceedance probability $p = 0.44$). As in Experiment 1, this indicates that participants learned only locally combined with sparse extrapolations. However, in making their decisions, they balanced exploration and exploitation more explicitly by directly weighting the expected rewards and the associated uncertainty of different arms. Compared to Experiment 1, more participants were described well by the a context-blind Kalman Filter learning strategy, combined with an Expected Improvement (mean AIC = 387, $n = 11$, $p = 0.28$). This may be due to the increased difficulty of the task, which made memorization of rewards for each context less feasible, and may have moved a number of participants to give up on learning the relation between contexts and rewards.

Marginalizing the protected probability of exceedance over all acquisition functions showed that the Kalman filter ($p = 0.53$) and the GP with an Orstein-Uhlenbeck kernel ($p = 0.47$) were the two mostly applied learning strategies

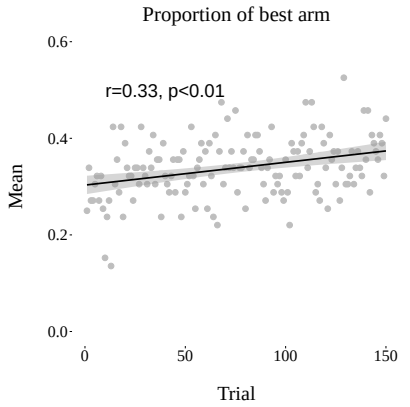
bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a) Overall average scores.



(b) Proportion of 4th arm.



(c) Proportion of best arm.

Figure 4. Average overall score (a), mean score per round (c), and proportion of best arm chosen per round (c) for the continuous-linear CMAB task.

overall. Again, this shows that most participants learned very locally only or relied on a context-blind strategy. Marginalizing over the learning models showed that the UCB is by far the best acquisition function ($p = 0.45$).

Experiment 3: Continuous-Non-Linear CMAB

The previous experiments showed that many people were able to learn how a general multi-feature context differentially affects the rewards associated to decision alternatives. The goal of the third experiment was to investigate whether this would still be the case in an even more complex situation where the contexts consist of continuous features which are non-linearly related to the rewards associated to the arms. In order to cover a wide range of non-linearities, the functions relating contexts to rewards were themselves sampled from a Gaussian process prior.

Participants

60 participants (28 female) with a mean age of 29 (SD=8.2) were recruited via Amazon Mechanical Turk and received $0.3 as a basic reward and a performance-dependent reward of up to $0.5.

Task and Procedure

The task was identical to that of Experiment 2, apart from the functions mapping inputs to outputs, which were drawn from a Gaussian process prior:

$$
\begin{aligned}
f_1(s_t) &= 50 + f_1(s_{1,t}, s_{2,t}) \\
f_2(s_t) &= 50 + f_2(s_{2,t}, s_{3,t}) \\
f_3(s_t) &= 50 + f_3(s_{3,t}, s_{1,t}) \\
f_4(s_t) &= 50 \\
f_j &\sim \mathcal{GP}(\mu, \Sigma), j = 1, \ldots, 3
\end{aligned}
$$

where the kernel used to sample the functions from was a Radial Basis kernel with a length-scale of $\theta_2 = 2$.

As in Experiment 2, the features were described numerically and could take values between -10 and 10. These values were determined at random and sampled from a uniform distribution $s_{i,t} \sim \mathcal{U}(-10, 10)$. As before, the average expectation for all planets was 50 and the variance for the fourth arm was the lowest.

The procedure was identical to the one of Experiment 2.

Results

Behavioral results. On average, participants obtained rewards of 55.35 (SD = 6.33), which is again above chance level, $t(59) = 5.85$, $p < 0.01$. 8 participants performed worse than expected by chance level and were therefore excluded from the following analysis.

Average scores increased significantly over trials, $r = 0.19$, $p < 0.01$ (see Figure 6b), although the correlation was smaller than in Experiment 2, which might be due to the increase in difficulty of the task.

The proportion of participants choosing the best option given the current context increased over trials, $r = 0.12$,

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

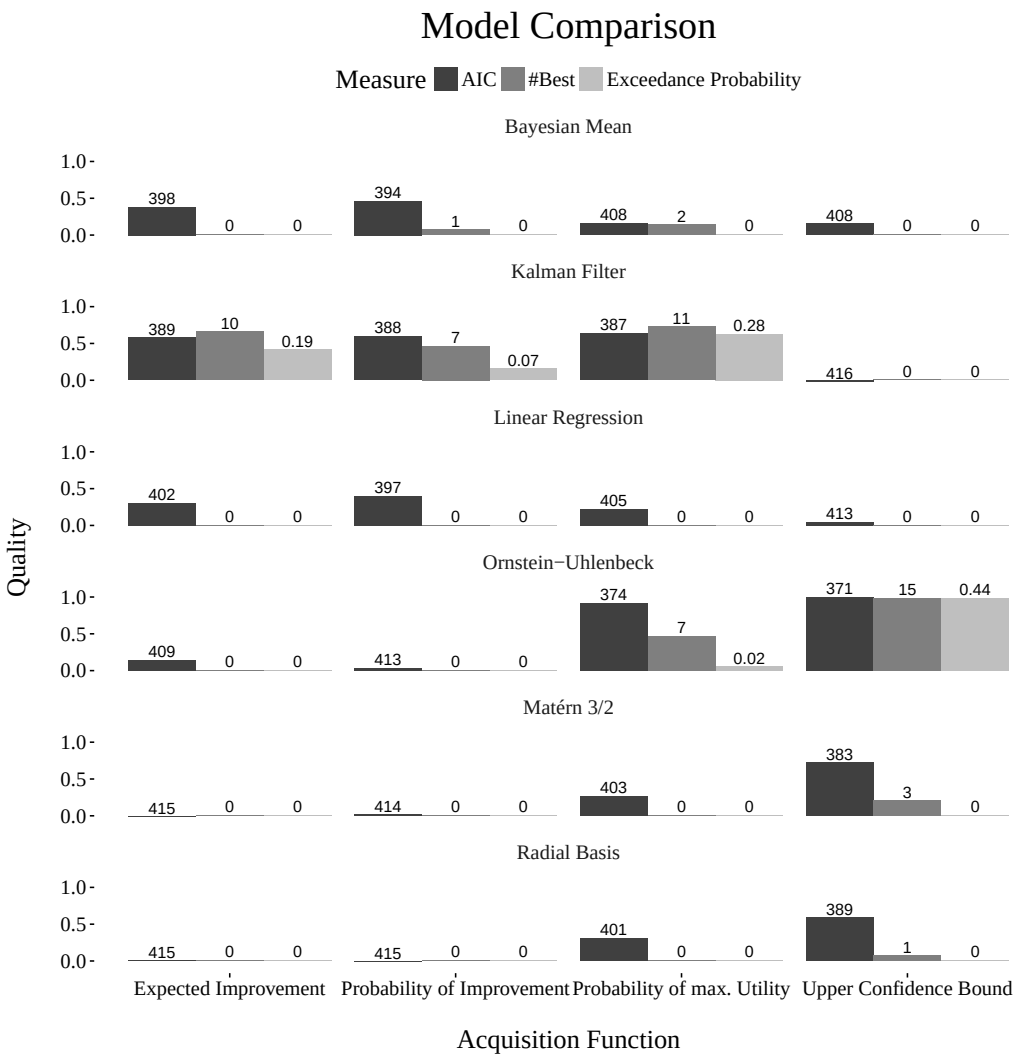


Figure 5. Results of model comparison for CMAB with continuous-linear cues when accounted for random model (mean AIC=415.89, 0 participants explained best, exceedance probability=0). Results were standardized to fit on one scale.

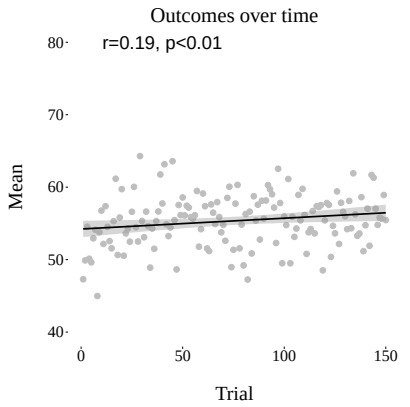
$p < 0.05$, but less marked than in the previous experiments (see Figure 6c). The proportion of choosing the non-contextual arm did again not significantly decrease over time, $r = 0.04$, $p > 0.05$. In a non-linear scenario, it might make even more sense for both contextual and context-free learning strategies to choose the 4th arm as it generates the mean output almost certainly.

Modelling results. Modelling results are shown in Figure 7. It can be seen that two models describe participants' behavior well. The best performing model combines the Ornstein-Uhlenbeck kernel with a Upper Confidence bound decision strategy. This model has a mean AIC of 377, describes $n = 10$ participants best, and has a protected probability of exceedance of $p = .69$. The next best model combines the context-free Kalman Filter with a Probability of maximum utility acquisition function. This model has a mean AIC of 379, also describes $n = 10$ participants best, but has a lower protected probability of exceedance of $p = 0.24$. As this last experiment required participants to learn three different non-linear functions, it might have been too difficult for some participants to learn the functions, so that they reversed back to learning in a purely context-free manner. Those who were able to learn contextually again only do so very locally and combined this with a decision strategy which explicitly trades off mean expectations and their uncertainties. Marginalizing over the acquisition functions or learning models in this scenarios shows the same results as the individual analysis above as there really were only two models describing participants' behaviour well.

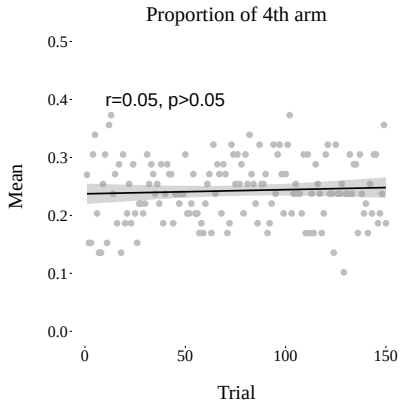
Discussion and Conclusion

We have introduced the contextual multi-armed bandit (CMAB) task as a paradigm to investigate behaviour in situa-

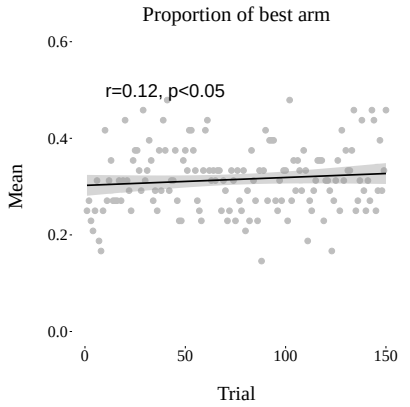
bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a) Overall average scores.



(b) Proportion of 4th arm.



(c) Proportion of best arm.

Figure 6. Average overall score (a), mean score per round (c), and proportion of best arm chosen per round (c) for the continuous-non-linear CMAB task.

tions where participants have to learn functions and simultaneously make decisions according to the predictions of those functions. The CMAB is a natural extension of past research on learning in multi-armed bandit tasks as well as research on function learning. In three experiments, we assessed people's performance in a CMAB task where a general context affected the rewards of options differently (i.e., each option had a different function relating contexts to rewards). Even though learning multiple functions simultaneously is likely to be more complex than learning a single function (as is common in previous studies on function learning and multiple cue probability learning), on average, participants were able to perform better than expected if they were unable to take the contexts into account. This was even the case in a rather complex situation where the functions were drawn from a general class of non-linear function, although performance dropped here compared to a simpler situation with linear functions.

In an environment where the contexts were defined by binary features, participants' learning appeared to be mostly locally-focused, relying on a universal learning strategy that matches a Gaussian process regression with an Ornstein-Uhlenbeck kernel. In basing their decisions on the learned functions, they appeared to rely on a strategy in which they focus on the expected improvement over past outcomes. In an environment with continuous contextual features which are linearly related to rewards, this picture changed somewhat; while participants again seemed to learn locally, here they appeared to trade-off their expectations and uncertainties more explicitly (best described by an Upper Confidence Bound acquisition function). In an environment involving non-linear functions sampled from a Gaussian Process prior, we again found that most learners were best-described as locally learning the functions and mapping their expectations to choices by explicitly balancing expected rewards and their uncertainties. However, this task was more difficult and a proportion of participants appeared unable to learn the functions, performing more in line with a context-blind learning strategy (Kalman filter) that treats the task as a restless bandit in which the expected rewards fluctuate over time but where these fluctuations are not predictable from the changes in context. The combination of a Kalman filter learning model with a "probability of maximum utility" decision strategy that described these participants best has been found to describe participants behaviour well in an actual restless bandit task Speekenbrink and Konstantinidis (2015).

Modelling function learning as Gaussian process regression allowed us to incorporate both rule-based and similarity-based learning in one framework. We found that most participants appeared to rely on an Ornstein-Uhlenbeck kernel, which assumes very unsmooth functions and generalizes only very locally from previous observations to similar contexts. This is close to a combination of exemplar learning, in which past observations are memorized and averaged when making new predictions, and sparse, similarity-based extrapolation. Such a memorization strategy is plausible when the number of unique contexts is relatively small. However, this kernel (as well as the Matérn 3/2 kernel, which assumes slightly smoother functions) also captured contextual

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

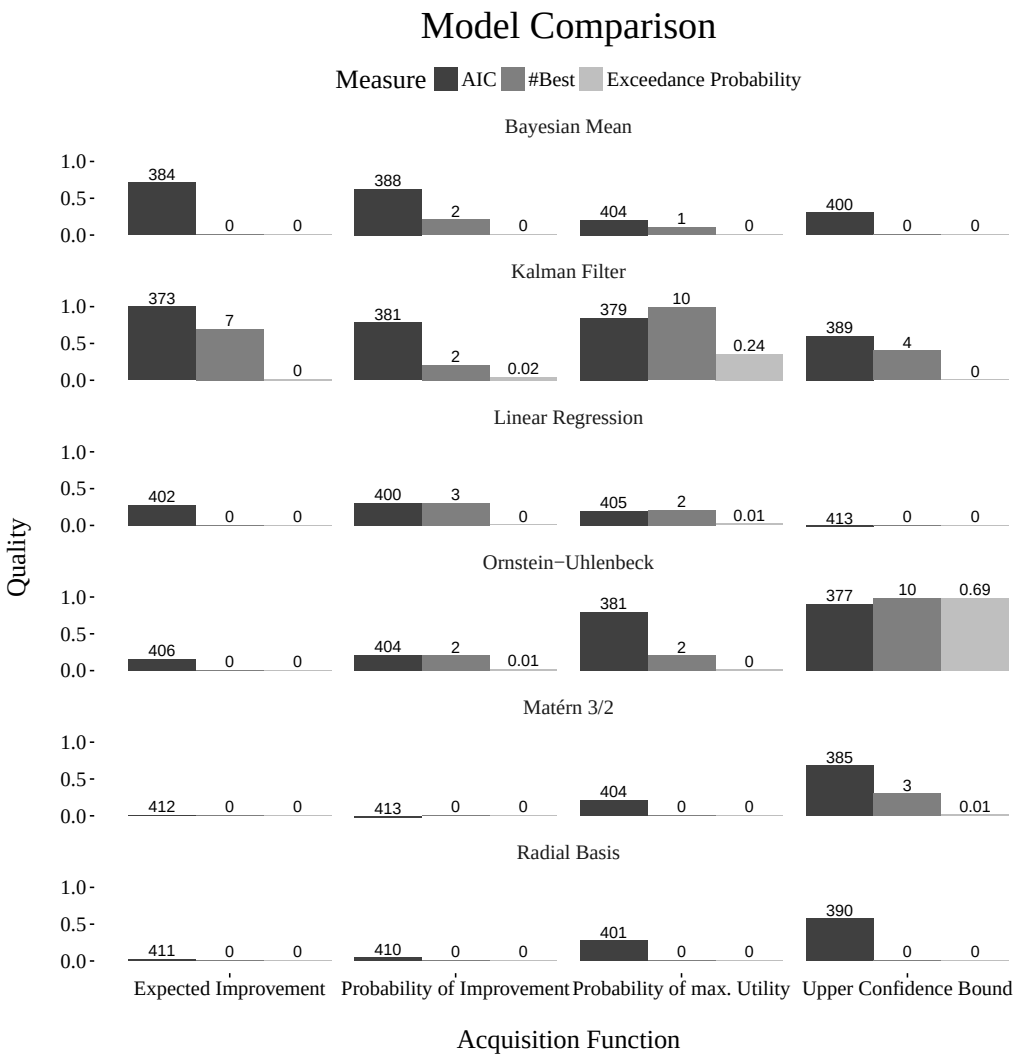


Figure 7. Results of model comparison for CMAB with continuous-non-linear cues when accounted for random model (mean AIC=415.89, 4 participants explained best, exceedance probability=0). Results were standardized to fit on one scale.

learners in experiments where the contexts were defined by continuous features and there were thus many more unique contexts. While previous research on function learning has found a strong bias towards linear functions in such environments (e.g., Lucas et al., 2015), we did not find such a bias in the present experiments. This could be due to the increased complexity of learning multiple functions simultaneously, or due to participants' learning the functions with the purpose of making good decisions, rather than to accurately predicting the outcomes as such. While good performance in standard function learning experiments requires accurate knowledge of a function in its whole domain, more course-grained knowledge may suffice in our CMAB task, where it suffices to know which function has the maximum output for a given context. Although the true functions relating contexts to rewards were smoother than those assumed by an Ornstein-Uhlenbeck kernel, approaching the task with a relatively unsmooth kernel may be wise when there is uncertainty about the level of smoothness in the environment. Assuming unsmooth functions and learning in a situation of smoother than expected functions will lead to smaller mismatched learning errors than the other way around, that is, expecting smooth functions and having to learn in unsmooth environments (see Sollich, 2001).

When making decisions based on their functional knowledge, participants appear to adapt their strategy to the task at hand. They seem to try and improve upon past outcomes within a relatively simple scenario with binary contextual features, while they trade off between expectations and their uncertainties more explicitly as the environment gets more complicated.

When there are few unique and distinct contexts, it might

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

be possible to memorize the average outcomes for those contexts, such that trying to maximally improve upon the current best option may be a feasible and efficient strategy. As the learning environment becomes more complex, a strategy that incorporates exploration more explicitly might be needed. Upper confidence bound sampling is not only the only acquisition function with provable good regret (Srinivas, Krause, Kakade, & Seeger, 2012), but it has also been proposed as a dynamic shaping bonus within the exploratory choice literature before (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006).

The present experiments focused on a general context which differentially affected the outcomes of options. Future studies utilizing the CMAB paradigm could incorporate contextual features which are option-specific (e.g., the type of restaurant) as well as general features (e.g., the area in which the restaurants are located), possibly allowing these to interact (e.g., a seafood restaurant might be preferable to a pizzeria in a fishing village, but not a mountain village).

To make the task more true to real-life decision situations, future research could adapt the reward functions to incorporate costs of taking actions or obtaining poor outcomes (see Schulz, Huys, Bach, Speekenbrink, & Krause, 2016). Research utilizing the CMAB paradigm also has the potential to be applied to more practical settings, for example military decision making, clinical gambling, or financial investment scenarios, to name just a few examples of decision making that normally involve both learning a function and making decisions based on expected outcomes. In general, we believe that incorporating context into models of reinforcement learning and decision making provides a fruitful avenue for future research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on Automatic Control*, *19*, 716–723. doi: 10.1109/TAC.1974.1100705

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*, 215–233. doi: 10.1002/bdm.443

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The iowa gambling task and the somatic marker hypothesis: Some questions and answers. *Trends in Cognitive Sciences*, *9*, 159–162. doi: 10.1016/j.tics.2005.02.002

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 38–50.

Bouton, M. E., & King, D. A. (1983). Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*, 248.

Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua*. Educational Testing Service.

Carroll, J. D., & De Soete, G. (1991). Toward a new paradigm for the study of multiattribute choice behavior: Spatial and discrete modeling of pairwise preferences. *American Psychologist*, *46*, 342.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 933–942. doi: 10.1098/rstb.2007.2098

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

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986. doi: 10.1037/0278-7393.23.4.968

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Comput Biol*, *11*, e1004567.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*, 197.

Gershman, S. J., Malmaud, J., Tenenbaum, J. B., & Gershman, S. (2016). *Structured representations of utility in combinatorial domains.*

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

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory*, *9*, 408–418.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. In *Advances in Neural Information Processing Systems* (pp. 553–560).

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*, 517–523.

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

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.

bioRxiv preprint doi: https://doi.org/10.1101/081091; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Krause, A., & Ong, C. S. (2011). Contextual Gaussian process bandit optimization. In *Advances in neural information processing systems* (pp. 2447–2455).

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, *86*, 97–106.

Laureiro-Martínez, D., Brusoni, S., & Zollo, M. (2010). The neuroscientific foundations of the exploration-exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, *3*, 95–115. doi: 10.1037/a0018495

Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes*, *116*, 286–295. doi: 10.1016/j.obhdp.2011.05.001

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).

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

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*, 61.

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, *13*, 2069–2106.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*, 24–42. doi: 10.3758/BF03196347

McDaniel, M. A., Dimperio, E., Griego, J. A., & Busemeyer, J. R. (2009). Predicting transfer performance: A comparison of competing function learning models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 173.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., … Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*, 191–215. doi: 10.1037/dec0000033

Mockus, J., Tiesis, V., & Zilinskas, A. (1978). The application of bayesian methods for seeking the extremum. *Towards global optimization*, *2*, 2.

Neimark, E. D., & Shuford, E. (1959). Comparison of predictions and estimates in a probability learning situation. *Journal of Experimental Psychology*, *57*, 294.

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2015). *Straight choices: The psychology of decision making* (2nd ed.). Hove, UK: Psychology Press.

Rasmussen, C. E. (2006). Gaussian processes for machine learning.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies – revisited. *Neuroimage*, *84*, 971–985.

Schulz, E., Huys, Q. J., Bach, D. R., Speekenbrink, M., & Krause, A. (2016). Better safe than sorry: Risky function exploitation through safe optimization. *arXiv preprint arXiv:1602.01052*.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2016). *Probing the compositionality of intuitive functions* (Tech. Rep.). Center for Brains, Minds and Machines (CBMM).

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., & Gershman, S. J. (2015). Assessing the perceived predictability of functions. *Proceedings of the 37th annual conference of the cognitive science society*, 2116–2121.

Sollich, P. (2001). Gaussian process regression with mismatched models. *arXiv preprint cond-mat/0106475*.

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 task. *Topics in Cognitive Science*, *7*, 351–367. doi: 10.1111/tops.12145

Speekenbrink, M., & Shanks, D. R. (2008). Through the looking glass: A dynamic lens model approach to multiple cue learning. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 409–429). Oxford University Press.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. W. (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *Information Theory, IEEE Transactions on*, *58*, 3250–3265.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*, 1004–1017.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179. doi: 10.1016/j.jmp.2008.11.002

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). MIT press Cambridge.

Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., & Harvey, N. (2016). Incorporating conflicting descriptions into decisions from experience. *Organizational Behavior and Human Decision Processes*, *135*, 55–69. doi: 10.1016/j.obhdp.2016.05.005