Generalization and search in risky environments

Eric Schulz

Harvard University

Charley M. Wu

Max Planck Institute for Human Development

Quentin J.M. Huys

ETH and University of Zürich

Andreas Krause

ETH Zürich

Maarten Speekenbrink

University College London

Author Note

Eric Schulz, Department of Psychology, Harvard University; Charley M. Wu, Center for Adaptive Rationality and Cognition, Max Planck Institute for Human Development; Quentin J.M. Huys, Translational Neuromodeling Unit, ETH Zürich and University of Zürich; Andreas Krause, Department of Computer Science, ETH Zürich; Maarten Speekenbrink, Department of Experimental Psychology, University College London

A preliminary version of this work has been published in the proceedings of the 38th Annual Meeting of the Cognitive Science Society as Schulz, Huys, Bach, Speekenbrink, and Krause (2016).

Correspondence concerning this article should be addressed to Eric Schulz, E-mail: ericschulz@fas.harvard.edu

Abstract

How do people pursue rewards in risky environments, where some outcomes should be avoided at all costs? We investigate participants' search for spatially correlated rewards in scenarios where one must avoid sampling rewards below a given threshold. This requires participants to not only balance exploration and exploitation, but also to reason about how to avoid potentially risky areas of the search space. Within risky versions of the spatially correlated multi-armed bandit task, we show that participants' behavior is aligned well with a Gaussian process function learning algorithm, which chooses points based on a safe optimization routine. Moreover, using leave-one-block-out cross-validation, we find that participants adapt their sampling behavior to the riskiness of the task, although the underlying function learning mechanism remains relatively unchanged. These results show that participants can adapt their search behavior to the adversity of the environment and enrich our understanding of adaptive behavior in the face of risk and uncertainty.

Keywords: Exploration-Exploitation; Generalization; Function learning; Risky choices

Generalization and search in risky environments

Introduction

Your phone rings. It is your parents. They are on their way for a surprise visit. You can hear the engine of their car running as you talk to them. They will arrive in a few hours and ask if you could prepare something for dinner. Your mum jokes that they already had beans on toast for lunch. Aiming to amaze them with a unique culinary experience, you decide to prepare something extraordinary, something they have never eaten before. As you open your fridge and kitchen cupboards, you find a plethora of ingredients at your disposal. In your head, you go through different combinations of ingredients, imagining how their taste combines and interacts to produce a—hopefully memorable—culinary experience. You have enough time to try out some combinations, experiencing the resulting taste, and thereby learning about the effects of unusual combinations and methods of preparation. While you can be adventurous, you need to be sure to avoid certain options; you cannot risk trying inedible, poisonous, or otherwise disastrous dishes.

This scenario is an example of a multi-armed bandit task (Robbins, 1985; Srinivas, Krause, Kakade, & Seeger, 2009; Steyvers, Lee, & Wagenmakers, 2009), where there are a number of options or "arms" of the bandit (e.g., the possible dishes) which lead to initially unknown and stochastic outcomes or rewards (e.g., the taste of the dish), that are related to a set of features (e.g., the ingredients, the method of preparation, and so forth). Through experience, you can learn a function which maps features to rewards, and use this knowledge to maximize the overall rewards gained over repeated plays of the bandit. A key challenge for optimal behavior in such tasks is framed by the exploration-exploitation dilemma (Gittins, 1979; Laureiro-Martínez, Brusoni, & Zollo, 2010): should you choose an option that you know will likely lead to a high reward, or try an unknown option to experience its outcome and thereby learn more about the function mapping features to rewards, increasing the chances of gaining even higher rewards in the future?

Single-mindedly focusing on optimizing outcomes is frequently ill-advised as there

might be further constraints which one has to take into account. For example, you may need to avoid a particularly bad outcome (e.g., poisonous food) at all cost. In order to satisfy this constraint, you should only explore options that—while uncertain—are likely to be "safe". Such restricted exploration-exploitation problems are indeed common in daily life, from choosing which restaurant to visit (avoid food poisoning), where to buy a second-hand car (avoid buying a lemon), to finding the shortest route home (avoid dangerous terrain). In our previous research on human behavior in contextual (Schulz, Konstantinidis, & Speekenbrink, 2017) and spatially-correlated multi-armed bandits (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2017a), we found that human reward seeking in the absence of constraints can be robustly described by a combination of a universal function learning mechanism and a decision strategy which explicitly balances an option's expected reward and its attached uncertainty. The function learning mechanism was formalized as Gaussian process regression – a form of non-parametric Bayesian regression that adapts its complexity to the data at hand (Griffiths, Lucas, Williams, & Kalish, 2009; Rasmussen, 2006). The decision strategy was formalized as upper confidence bound sampling (UCB; Auer, 2002) strategy.

In the present study, we expand on our previous work by introducing scenarios with additional constraints: unsafe options —defined as options which produce outputs below a given threshold— which should be avoided at all costs. We assess how people behave when they have to maximize accumulated rewards whilst avoiding (momentary) outcomes below the threshold. The task is presented as a spatially correlated multi-armed bandit in which participants choose an input, and then observe and accrue the output of an underlying function which maps spatial locations to expected rewards. In two experiments with a uni- and bivariate spatially-correlated multi-armed bandit, we find that participants adapt their exploration-exploitation strategy to the additional constraints of risky situations, but utilize the same underlying learning mechanism.

From a computational perspective[1], the task of maximizing rewards while

[1]Marr (1982) famously proposed to analyze intelligent systems on three different levels: the computational (what is the task the system is trying to solve), the algorithmic level (how does it solve it), and the implementation level (how is the solution implemented).

avoiding unsafe inputs can best be solved by a safe optimization algorithm (Sui, Gotovos, Burdick, & Krause, 2015). This algorithm tries to find points that are likely to be safe, and then expands on the set of safe points while also attempting to optimize the underlying function. When analyzing how humans choose from the three sets of points identified by a safe optimization algorithm (i.e., safe, expanding, and maximizing points), we find that choices are strongly influenced by a tendency to stay safe.

From an algorithmic perspective, it is possible to predict individual participant choices by combining different models of learning with multiple decision strategies, and then perform model comparison using out-of-sample prediction accuracy. Whereas the estimated parameters of the learning model remain relatively unchanged, we find that participants seem to adapt their decision strategy to whether or not they need to avoid unsafe outcomes, as predicted by a decision strategy that focuses on staying safe.

Our results point to the relevance of safe reinforcement learning (Berkenkamp, Turchetta, Schoellig, & Krause, 2017) for explaining human behavior in naturalistic tasks and enrich our notion of how people strategically adapt their behavior to the risk and uncertainty of the environment. In particular, whereas the way in which people generalize over different options remains similar across different riskiness conditions, they tend to adapt their sampling strategy by focusing more on safe actions as the situations become more risky.

General task description

We use a variant of the spatially correlated multi-armed bandit (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2017b), where the rewards of each option (i.e., arm) are correlated according to how close they are to each other. Intuitively, nearby arms tend to have similar rewards, with the level of similarity decreasing over larger distances. The options are either univariate input values placed along a line, or bivariate input values placed on a grid. Each discretized input value represents a playable arm of the bandit. Traditionally, the goal in such tasks is to maximize cumulative payoffs by sequentially choosing one of the N-arms of the bandit that stochastically generate

rewards (Lai & Robbins, 1985; Steyvers et al., 2009), with learning happening independently for each arm (e.g., through associative learning). In our case, because proximate arms generate similar rewards due to the spatial correlations, there is the opportunity to form inductive beliefs about the rewards of untried options by learning the function that maps the spatial location of options to their rewards. This task allows us to study how people generalize their experience to generate beliefs about novel options, and how this process influences their search behavior (Wu et al., 2017a).

Importantly, we add a constraint such that participants need to avoid rewards below a given threshold. If participants obtain rewards above the threshold, they collect the reward and continue to the next trial where they are again asked to choose an input. If they obtain a reward below the threshold, they forego the reward, end their round, and lose the opportunity to collect further rewards within the current round. The additional requirement of avoiding unsafe options in the Risky Spatially Correlated Bandit makes generalization even more important, as it now serves to identify not only highly rewarding options, but also unsafe options. In contrast to unconstrained spatially correlated multi-armed bandits, where good performance does not require accurate knowledge of the function in regions of low rewards, our risky version requires people to learn about both regions of low and high reward.

Function learning as model of generalization

We assume that generalization within spatially correlated multi-armed bandits can be described as a function learning mechanism that learns a function mapping the spatial context of each arm to expectations of reward. We use Gaussian process regression (Rasmussen, 2006; Schulz, Speekenbrink, & Krause, 2016) as an expressive model of human function learning. Gaussian process regression is a non-parametric Bayesian approach towards function learning which can perform generalization by making inductive inferences about unobserved outcomes. In past research we found that Gaussian process regression captures the inductive biases of human participants in a variety of explicit function learning tasks (Schulz, Tenenbaum, Duvenaud,

Speekenbrink, & Gershman, 2016), and provides an accurate description of human generalization in contextual and spatially correlated multi-armed bandits without the presence of unsafe outcomes (Schulz et al., 2017; Wu et al., 2017a).

Gaussian process regression integrates both rule-based and similarity-based approaches towards function learning and has originally been proposed as a rational model of human function learning by Lucas, Griffiths, Williams, and Kalish (2015). Here, we use Gaussian process regression both as a rational model, and as a component in our models which describe behavior on an algorithmic level. Thus, we use an approach that bridges the gap between two levels of descriptions (Griffiths, Lieder, & Goodman, 2015; Griffiths, Vul, & Sanborn, 2012).

To categorize participants' choices from a computational level, we assess the correspondence between input points preferred by a Gaussian process-based safe optimization algorithm and those preferred by the participants in our experiments. To determine the model which describes behavior on the algorithmic level, we combine Gaussian process regression with different decision strategies, some of which are risk-averse and some of which are not. We then use participant-wise cross validation to compare the resulting models with other models that do not apply generalization.

Gaussian process function learning

A Gaussian process defines a distribution over functions (see Rasmussen, 2006; Schulz, Speekenbrink, & Krause, 2016, for an introduction). Let $f : \mathcal{X} \mapsto \mathbb{R}$ denote a function over input space $\mathcal{X}$ (i.e., options or arms) that maps to real-valued scalar outputs (i.e., rewards). The function is assumed to be a random draw from a Gaussian process:

$$f \sim \mathrm{GP}(m, k), \tag{1}$$

where m is a mean function specifying the expected output of the function given input $\mathbf{x}$, and k is a kernel (or covariance) function specifying the covariance between outputs:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{2}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right]. \tag{3}$$

Intuitively, the kernel encodes an inductive bias about the function's expected smoothness. We follow standard conventions and set $m(x) = 0$.

Conditional on observed data $\mathcal{D}_t = \{\mathbf{x}_j, y_j\}_{j=1}^t$, where $y_j \sim \mathcal{N}(f(\mathbf{x}_j), \sigma_\epsilon^2)$ is a noise-corrupted draw from the underlying function (σ_ϵ^2 is the noise variance), the posterior predictive distribution of the function value for a new input $\mathbf{x}_*$ is Gaussian with mean and variance given by:

$$\mathbb{E}[f(\mathbf{x}_*)|\mathcal{D}_t] = \mathbf{k}_{t*}^\top(\mathbf{K}_t + \sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{y}_t \tag{4}$$

$$\mathbb{V}[f(\mathbf{x}_*)|\mathcal{D}_t] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{t*}^\top(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}\mathbf{k}_{t*}, \tag{5}$$

where $\mathbf{y}_t = [y_1, \ldots, y_t]^\top$, $\mathbf{K}_t$ is the $t \times t$ matrix of covariances evaluated at each pair of observed inputs, and $\mathbf{k}_{t*} = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_t, \mathbf{x}_*)]$ is the covariance between each observed input and the new input $\mathbf{x}_*$.

A common choice of kernel function is the radial basis function kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{\lambda}\right), \tag{6}$$

where the length-scale λ governs how quickly correlations between points $\mathbf{x}$ and $\mathbf{x}'$ decay towards zero as their distance increases[2].

[2]Sometimes the RBF kernel is specified as $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2l^2}\right)$ whereas we use $\lambda = 2l^2$ for simplicity.

Rational model

Given a learned representation of a function at time t, this knowledge can be used to choose the next input at time $t+1$ in a close to rational way. This is done through a *decision strategy* that takes the predicted mean $\mu(\mathbf{x})$ and uncertainty $\sigma(\mathbf{x})$ for each input, and produces a criterion governing which input to choose next in order to balance exploration and exploitation (Brochu, Cora, & De Freitas, 2010; Schulz, Speekenbrink, & Krause, 2016).

A strategy that can cope with the additional requirement to avoid outcomes below a threshold in a close-to-rational way, is the *safe optimization algorithm* put forward by Sui et al. (2015). This algorithm uses Gaussian process regression to form beliefs about different arms' predictive payoff distributions at time point t. It first defines a *safe set* of possible inputs $\mathcal{S}_t$ that are likely to provide outputs above the threshold, and then further separates the safe set into a set of *maximizers* ($\mathcal{M}_t$, inputs that are likely to provide the maximum output) and *expanders* ($\mathcal{G}_t$, inputs that are likely to expand the safe set). Following Berkenkamp and Schoellig (2015), we define the upper and lower bounds of a confidence interval by adding the current expectation of reward $\mu_{t-1}(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})|\mathcal{D}_{t-1}]$ and the estimated uncertainty $\sigma_{t-1}(\mathbf{x}) = \sqrt{\mathbb{V}[f(\mathbf{x})|\mathcal{D}_{t-1}]}$:

$$u_t(\mathbf{x}) = \mu_{t-1}(\mathbf{x}) + \omega\sigma_{t-1}(\mathbf{x}) \tag{7}$$

$$l_t(\mathbf{x}) = \mu_{t-1}(\mathbf{x}) - \omega\sigma_{t-1}(\mathbf{x}). \tag{8}$$

The parameter ω determines the width of the confidence bound, and we set it to $\omega = 3$ to assure high safety for the rational safe optimization algorithm. Using these bounds, we can define the safe set $\mathcal{S}_t$ as all the input points in the set of available inputs $\mathcal{X}$ that are likely to lead to output values above the safe threshold $h_{\min}$,

$$\mathcal{S}_t = \{\mathbf{x} \in \mathcal{X} | l_t(\mathbf{x}) \geq h_{\min}\} \tag{9}$$

This means that points are considered to be safe if their lower confidence bound is

above the provided threshold. This is intuitive as one would expect the output for these points to rarely fall below the threshold.

The set of potential maximizers $\mathcal{M}_t$ contains all safe inputs that are likely to obtain high output values; these are the safe inputs for which the upper confidence bound u_t is above the best current lower bound:

$$\mathcal{M}_t = \{\mathbf{x} \in \mathcal{S}_t | u_t(\mathbf{x}) \geq \max_{\mathbf{x}' \in \mathcal{X}} l_t(\mathbf{x}')\} \tag{10}$$

where $\mathbf{x}'$ is the best revealed reward at time t. This means that maximizers are input points that are expected to be at least as good as the best overall outcome in a worst case scenario.

To find the set of expanders, we first define

$$g_t(\mathbf{x}) = |\{\mathbf{x}' \in \mathcal{X} \setminus \mathcal{S}_t | l_{t,(\mathbf{x},u_t(\mathbf{x}))}(\mathbf{x}') \geq h_{\min}\}| \tag{11}$$

where $l_{t,(\mathbf{x},u_t(\mathbf{x}))}(\mathbf{x}')$ is the lower bound of $\mathbf{x}'$ based on past data and a optimistic outcome for $\mathbf{x}$ which provides a new upper bound $u_t(\mathbf{x})$. Put differently, this function is used to determine how many inputs might possibly be added to the safe set after choosing input $\mathbf{x}$ and observing the output it provides. This function is positive only if the new data point has a non-negligible chance to expand the safe set. The set of potential expanders is then defined as

$$\mathcal{G}_t = \{\mathbf{x} \in \mathcal{S}_t | g_t(\mathbf{x}) \geq 0\} \tag{12}$$

This expander set is assessed by forward simulation and simply checks if the safe set is expected to be expanded (i.e., that more points will be in this set) by choosing a given input point.

For categorizing participants' decisions, we assess how much—if at all—participants' choice behavior is guided by options being safe, maximizers, and/or expanders. More precisely, we check if membership of an input point within the three

sets makes it more likely to be chosen.

In order to assess if set membership is related to participants' decisions, we will use mixed-effects logistic regression to determine the extent to which factors of the safe optimization algorithm influence their choices. The dependent variable in this analysis is, for each option, whether it was chosen or not by a given participant on a given trial. The predictors are indicator variables for an option's membership of the safe, the maximization, and the expander set. This analysis allows us to judge whether (i) a Gaussian process function learning model with parameters set to match the underlying task, combined with (ii) class membership specified by the safe optimization algorithm, can describe whether or not a participant chose an option on a given trial. This will be the first part of our analysis.

Models of learning and decision making

In the second part of our analysis, we make out-of-sample predictions about individual choices, using a combination of different learning models with multiple decision strategies. We first contrast two different models of learning before describing the decision strategies. The *function learning* model learns about the underlying value function relating the spatial locations of options to their expected rewards. The *option learning* model does not learn about an underlying function, but rather learns about each input individually by associating inputs with previously generated rewards.

Function learning model

For the function learning model, we use Gaussian process regression combined with a radial basis function kernel (Eq. 6). Using a radial basis function to model the extend of generalization across space is similar to Shepard et al. (1987)'s proposal of a universal law of generalization and has previously been implemented in a non-Bayesian model of function learning by Busemeyer, Byun, Delosh, and McDaniel (1997).

Option learning model

The option learning model uses a simple mean tracking approach to learn the distribution of rewards of each input individually. We implement a version which assumes rewards are normally distributed with a known variance (σ_ϵ^2) but unknown mean θ_j where the prior distribution of the mean is again a normal distribution. This implies that the posterior distribution for each mean is also a normal distribution:

$$p(\theta_j|\mathcal{D}_t) = \mathcal{N}(\mu_{j,t}, \sigma_{j,t}) \tag{13}$$

The mean $\mu_{j,t}$ and variance $\sigma_{j,t}^2$ of the posterior distribution for option j are only updated when that option is selected at trial t:

$$\mu_{j,t} = \mu_{j,t-1} + \delta_{j,t} G_{j,t} \left[y_t - \mu_{j,t-1}\right] \tag{14}$$

$$\sigma_{j,t}^2 = \left[1 - \delta_{j,t} G_{j,t}\right] \sigma_{j,t-1}^2 \tag{15}$$

where $\delta_{j,t} = 1$ if option j was chosen on trial t, and 0 otherwise. Intuitively, the estimated mean of the chosen option $\mu_{j,t}$ is updated based on the difference between the observed value y_t and the expected mean $\mu_{j,t-1}$, multiplied by $G_{j,t}$. At the same time, the estimated variance $\sigma_{j,t}^2$ is reduced by a factor of $1 - G_{j,t}$, where $G_{j,t}$ is defined as:

$$G_{j,t} = \frac{\sigma_{j,t-1}^2}{\sigma_{j,t-1}^2 + \sigma_\epsilon^2}, \tag{16}$$

σ_ϵ^2 is the error variance, which is estimated as a free parameter. We set the prior mean to the median value of the payoffs and the prior variance $\sigma_{j,0}^2 = 5$

This model does not generalize over unseen arms at all, but rather only learns locally about the distribution of rewards for each option separately (Wu et al., 2017a). We use it as a benchmark for comparisons in our cross-validation procedure. If this model predicts participants' behavior well, this means that participants do not generalize using the spatial structure of the environment, but rather learn about each

option independently, as is the case in a traditional multi-armed bandit.

Decision strategies

The learning models produce predictions about the distribution of rewards for each option in the search space, whereby we use a decision strategy to determine useful actions and predict choices. We compare four different strategies, two of which are designed for safe search and two of which are designed for risky search. Each strategy is based on a acquisition function, which determines a utility value for each option, with a softmax choice used to make probabilistic predictions about choice behavior.

Decision strategies for safe tasks. *Upper confidence bound sampling* directly trades off between the expected rewards and uncertainty. Given the posterior mean $\mu_{t-1}(\mathbf{x})$ and its attached standard error $\sigma_{t-1}(\mathbf{x})$, we calculate the acquisition function of the upper confidence bound as

$$\mathrm{UCB}_t(\mathbf{x}) = \mu_{t-1}(\mathbf{x}) + \beta\sigma_{t-1}(\mathbf{x}), \tag{17}$$

where the exploration factor β determines how much reduction of uncertainty is valued (relative to exploiting known high-value options) and is estimated as a free parameter. This sampling strategy has recently been found to describe human behavior well across different function exploration-exploitation tasks without additional constraints (Schulz et al., 2017; Wu et al., 2017a). Additionally, it has known performance guarantees in function optimization scenarios (Srinivas, Krause, Kakade, & Seeger, 2012). We will use it as a candidate for unconstrained function optimization tasks.

The *probability of improvement* (POI) strategy evaluates an option based on how likely it will be better than the best outcome ($\mathbf{x}^+$) observed so far:

$$\begin{aligned}\mathrm{POI}_t(\mathbf{x}) &= P\left(f(\mathbf{x}) \geq f(\mathbf{x}^+)\middle|\, \mathcal{D}_{t-1}\right) \\ &= \Phi\left(\frac{\mu_{t-1}(\mathbf{x}) - \mu_{t-1}(\mathbf{x}^+)}{\sigma_{t-1}(\mathbf{x})}\right)\end{aligned} \tag{18}$$

where $\Phi(\cdot)$ is the normal CDF. This rule calculates the probability for each option to

lead to an outcome higher than the option that has currently been observed (Kushner, 1964) and has recently been used in experiments involving multi-attribute choices (Gershman, Malmaud, & Tenenbaum, 2017).

Decision strategies for risky tasks. To define possible search strategies for risky situations, we consider two modifications of the decision strategies defined above. The *probability of being safe* (POS) is similar to the POI strategy, but assess the probability that a candidate input provides a reward above the safe threshold. Formally, if the threshold is $h_{\min}$, POS is defined as:

$$\begin{aligned} \mathrm{POS}_t(\mathbf{x}) &= P\left(f(\mathbf{x}) \geq h \middle| \mathcal{D}_{t-1}\right) \\ &= \Phi\left(\frac{\mu_{t-1}(\mathbf{x}) - h_{\min}}{\sigma(\mathbf{x})}\right). \end{aligned} \tag{19}$$

Instead of sampling by the probability to improve upon the best seen point so far, this sampling strategy only cares about maximizing the probability of being safe (i.e., sampling above the threshold).

Instead of valuing uncertainty positively, the *lower confidence bound algorithm* (LCB)

$$\mathrm{LCB}_t(\mathbf{x}) = \mu_{t-1}(\mathbf{x}) - \beta\sigma_{t-1}(\mathbf{x}), \tag{20}$$

tries to avoid highly uncertain options. As input areas with high uncertainty can also lead to possibly bad outcomes, this sampling strategy can be seen as a highly risk-averse but possibly not very adaptive approach to risky environments.

The difference between UCB and LCB corresponds to differences observed in risk-sensitive reinforcement learning when outcomes are either positive or negative (Niv, Edlund, Dayan, & O'Doherty, 2012).

Estimation and model comparison

For model fitting and evaluation, we use a cross-validation procedure in which we estimate parameters by maximum likelihood from a subset of the data and use these

estimated models to make out-of-sample predictions as a measure of predictive accuracy. A softmax function,

$$p_t(\mathbf{x}) = \frac{\exp(q_t(\mathbf{x})/\tau)}{\sum_{j=1}^{N} \exp(q_t(\mathbf{x}_j)/\tau)}, \tag{21}$$

transforms each model's criterion into a probability distribution over options, where $q_t(\mathbf{x})$ is the value of an option $\mathbf{x}$ according to the models' valuation of each option, and τ is the temperature parameter (i.e., lower values of τ indicate more precise predictions).

For the function learning model, we estimate λ (length-scale), for the option learning model σ_ϵ^2 (error variance), and for the upper and lower confidence bound sampling strategies β (exploration bonus for UCB, safety bonus for LCB). Additionally, all models include τ as a free parameter. We fit all models separately for each participant by cross-validated maximum likelihood estimation, using differential evolution algorithm (Mullen, Ardia, Gil, Windover, & Cline, 2009). Parameter estimates are constrained to positive values in the range $[\exp(-5), \exp(5)]$.

Cross-validation is performed for the safe and risky function exploration-exploitation objectives separately. Within all rounds, we use leave-one-block-out cross-validation to iteratively form a training set by leaving out a single round, computing a maximum likelihood estimate on the training set, and then generating out-of-sample predictions on the remaining round. This is repeated for all combinations of training and test sets, and for every participant individually. The prediction error (computed as log loss) is summed up over all trials, and is reported as average *predictive accuracy*, using a pseudo-R^2 measure that compares the total log loss prediction error for each model to that of a random model

$$R^2 = 1 - \log \mathcal{L}(\mathcal{M}_n) / \log \mathcal{L}(\mathcal{M}_{\text{rand}}), \tag{22}$$

where $\log \mathcal{L}(\mathcal{M}_n)$ is the log loss (negative log likelihood) of model n and $\log \mathcal{L}(\mathcal{M}_{\text{rand}})$ the log loss of a random model (which chooses options with equal probability). Intuitively, a $R^2 = 0$ corresponds to prediction accuracy equivalent to chance, while

$R^2 = 1$ corresponds to a perfect prediction accuracy.

Experiment 1: Univariate inputs

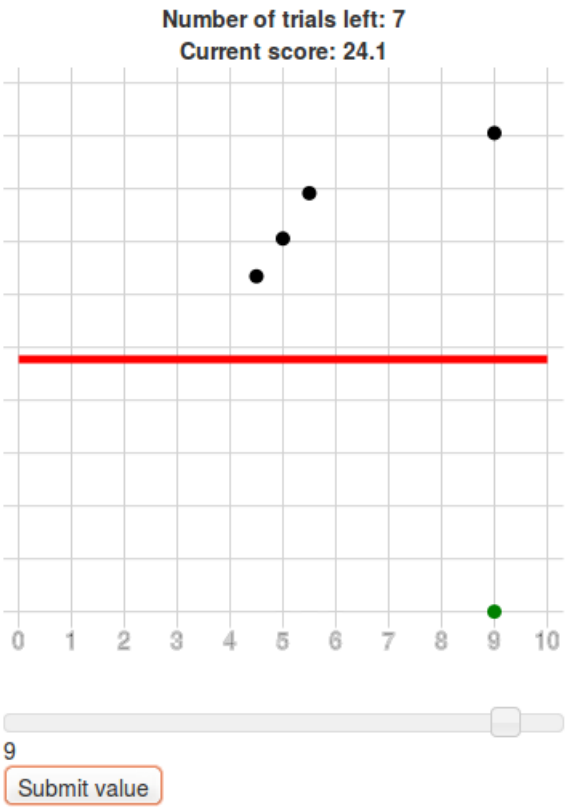

Figure 1. Screenshot of Experiment 1. The red line marks the safe threshold (outcomes below this threshold would immediately stop the round). Dots above the red line show observed outputs. The input can be chosen by moving the slider (the green dot shows the value) and selected by clicking "submit value".

The first experiment required participants to maximize unknown univariate functions $f\colon x \mapsto y$ by choosing discretized input values $x \in (0, 0.5, 1, \ldots, 10)$. This scenario is similar to a multi-armed bandit task (with $n = 21$ arms) in which all arms are ordered horizontally and where the outputs of the arms are correlated as a function of their distance. Additionally, we introduced the constraint that participants should avoid choosing options with a reward below the horizontal red line, or else forfeit the remaining trials in the round.

Participants

61 participants (36 female) with an average age of 32.95 (SD = 8.02) were recruited via Amazon Mechanical Turk and received $1 for their participation and a bonus of up to $1, in proportion to their overall score. The experiment took on average 12 minutes to complete.

Procedure

Participants were told they had to maximize outputs of an unknown function, while at the same time trying to avoid obtaining outputs below a given red line. After reading the instructions and performing an example round, they had to correctly answer 4 comprehension questions to check if they understood the instructions. There were 9 rounds in total and each round contained (at most) 10 trials. At the start of each round, participants were shown the output of a pre-selected input value. This input was randomly drawn from all inputs with outputs above the threshold.

On each round, participants had to choose an input with an output above the red line, as choosing a different input would end the current round and forfeit any additional rewards they could have earned by continuing selecting "safe" inputs on the given round. On each trial $t = 1, \ldots, 10$ in a block, they could choose an input value $x \in \{0, 0.5, 1, \ldots, 10\}$ to observe (and accrue) an output $y = f(x) + \epsilon$ with noise term $\epsilon \sim \mathcal{N}(0, 1)$. The underlying functions were sampled from a GP prior with a radial basis function kernel (length-scale λ=1). The objective was to maximize the sum of the obtained outputs over all trials in a round. A threshold $h_{\min}$ was introduced and a round was ended abruptly if an output below this threshold was obtained.

Before the first trial, an initial safe point above the threshold was sampled at random and provided to participants. A screenshot is shown in Figure 1. Outputs were scaled to be between 0 and 10 but such that the underlying maximum was never actually 10 in order to make the maximum not easily guessable. This was done by sampling a random number between 9 and 10 and using this number as the overall maximum for rescaling.

In order to see if different levels of risk influence participants' learning and sampling behavior, we manipulated the risk of obtaining outputs below the red line as a between-group factor, resulting into 6 groups for which the probability of sampling below the line was set to $p = [0.55, 0.6, 0.65, \cdots, 0.8]$. This means that, unknown to participants and before the start of each round, the red line was placed such that $p \times 100\%$ of the input points would produce an output lower than the red line. Participants received \$1 for participating in the experiment and a bonus of \$0.09 for every point acquired.

Behavioral results

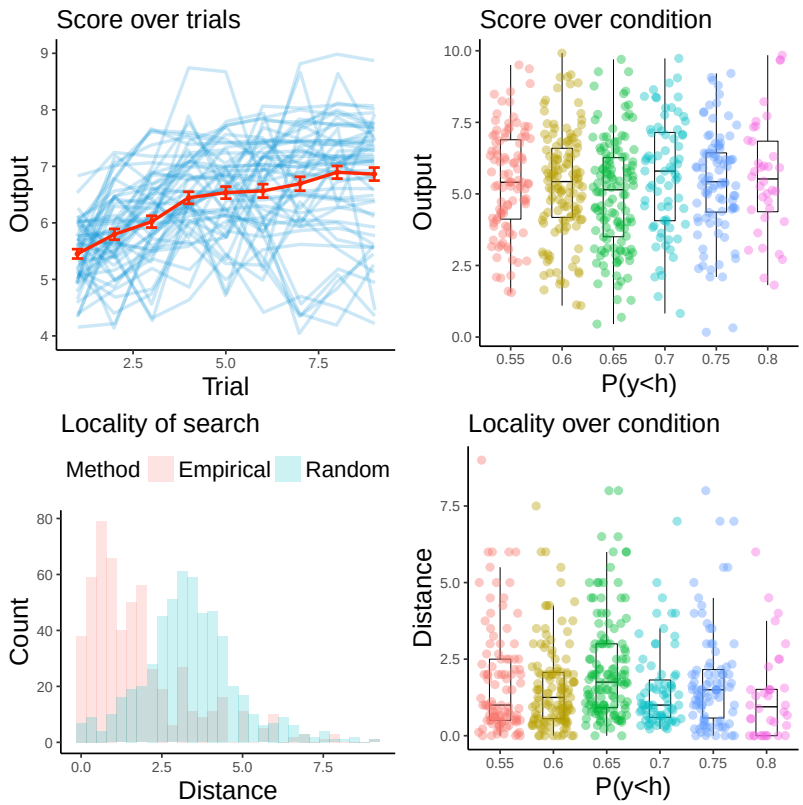

Figure 2. Results of Experiment 1. The upper left panel shows participants' scores over trials, including the average score (red) and standard errors in error bars. The upper right panel shows the average scores for the different risk conditions. The lower left panel shows the locality of chosen inputs as compared to a random sampler. The lower right panel shows the locality of chosen inputs by different riskiness-conditions.

Figure 2 shows the results of Experiment 1. In general, participants performed better than chance (mean score = 6.21, $t(61) = 12.32$, $p < .001$, $d = 1.57$) and improved

over time (mean correlation between trials and score $r = .2$, $t(60) = 6.87$, $p < .001$, $d = 0.88$). In addition, the average number of trials per block statistically exceeded what would be expected if participants chose completely at random ($t(60) = 22.69$, $p < .01$, $d = 2.9$), indicating that participants were successful at finding reward and avoiding risky options. Participants of different risk conditions did not perform significantly different from another (correlation between mean score and risk level: $r = .06$, $t(59) = 0.48$, $p > .6$). Participants also showed localized sampling behavior, choosing inputs more locally than a random sampling model ($t(60) = -22.1$, $p < .001$, $d = -2.83$), although participants in higher risk conditions did not choose more locally (correlation between risk level and average distance of consecutive inputs: $r = -.08$, $t(59) = -0.64$, $p = .52$). Therefore, participants learned within the task but were seemingly uninfluenced by the riskiness of the threshold.

Categorization of decisions

We used mixed-effects logistic regression analysis to determine the factors influencing participants' decisions. The dependent variable was whether each input was chosen or not on each trial for each participant. As predictors, we used indicator variables for an input's membership of the safe, maximization, and expander sets. Results indicated that the most plausible model was one that contains all variables as fixed effects and a participant-specific random intercept, indicating that participants were influenced by set membership in an overall similar fashion. The coefficients of the fixed effects are presented in Table 1 below.

Table 1
Results of the mixed effects logistic regression for Experiment 1.

	Estimate	S.E.	z value	Pr(>|z|)
Intercept	-3.71	0.12	-29.75	0.00
Maximizer	0.82	0.23	3.50	0.00
Expander	-0.44	0.20	-2.20	0.03
Safe	0.85	0.24	3.46	0.00

Comparing the magnitude of the slopes of the predictors, we can conclude that while all of the sets relate to participants' behavior, participants were mostly influenced

by whether or not a point was safe (Estimate = 0.85) or a maximizer (Estimate = 0.82). Being within the set of possible expanders was negatively related to whether or not participants would choose a given point (Estimate = -0.44). This shows that participants' behavior corresponded, at least to some extent, to the predictions generated by the rational model. However, they seemed to focus more on staying safe and maximizing locally rather than expanding the safe set. We next assessed if we could predict trial-by-trial decision behavior with more process-level modeling.

Trial-by-trial models

Assessing the trial-by-trial modeling results, we found that the overall best performing models were the Gaussian process model with a probability of being safe decision strategy as well as the option learning model with the same decision strategy.

Comparing the learning models by aggregating over decision strategies, we found that the option learning model outperformed the Gaussian Process learning model ($t(60) - 2.34$, $p < .05$, $d = 0.30$). This mirrors the behavioral finding that participants explored in a local manner, thus seemingly not generalizing much over different inputs and observed outcomes.

Comparing the decision strategies by aggregating over learning strategies, we found that the probability of being safe strategy predicted participants' behavior better than any of the other strategies, no matter whether it was combined with the option learning ($t(60) = 5.72$, $p < 0.001$, $d = 0.73$) or the Gaussian process function learning model ($t(60) = 6.28$, $p < 0.001$, $d = 0.80$).

Finally, there was no significant difference between the Gaussian process function learning and the option learning when both were combined with the probability of being safe sampling strategy ($t(60) = -0.71$, $p = 0.48$, $d = 0.09$).

We extracted the median parameter estimates of the Gaussian process learning model combined with the probability of being safe sampling strategy for each participant to check if they meaningfully tracked behavioral differences in the task (Figure 4). Overall, the model's predictions were relatively precise as indicated by low

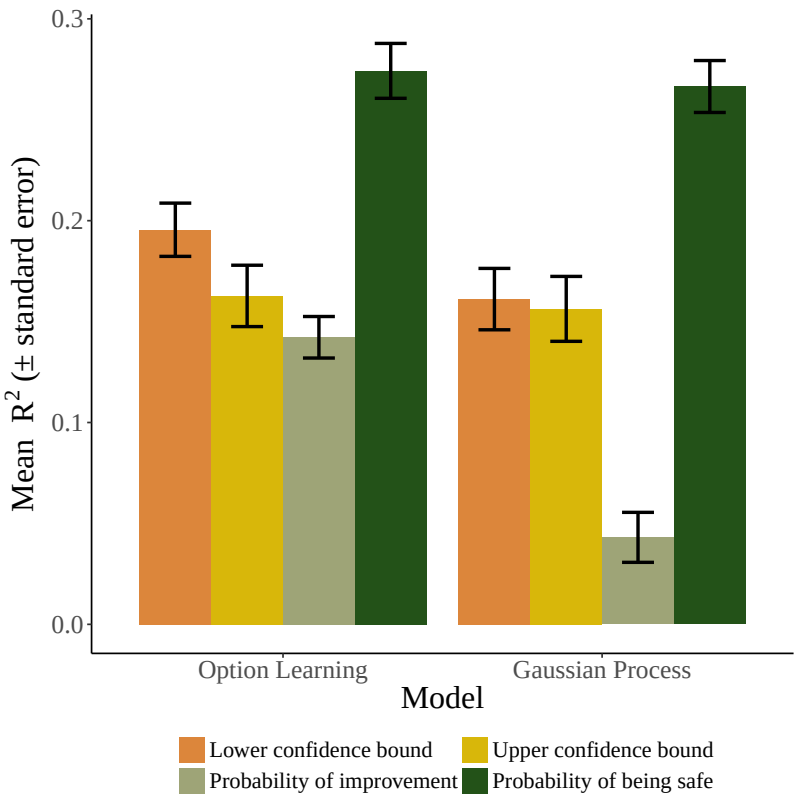

Figure 3. Results of the trial-by-trial learning models in Experiment 1.

estimates of the softmax temperature parameter τ (median estimate: $\hat{\tau} = 0.097$). Furthermore, the length-scale parameter of the Gaussian process indicated that participants seemed to somewhat generalize over different arms (median estimate $\hat{\lambda} = 4.43$; compared to the ground truth of $\lambda = 1$). Interestingly, people who generalized more performed worse overall ($r = -.36$, $t(59) = -2.99$, $p < .004$). This is most likely the result of the dual objectives participants were facing, which required them to generalize beyond encountered examples but to also sample safe options, which frequently required sampling rather locally.

Lastly, we used the participant-wise parameters of the Gaussian process function learning model paired with the probability of being safe sampling strategy (henceforth GP-POS) to simulate data within the task. This means that we let the model parameterized by participants' estimates of both λ and τ perform the exact same task as participants for the exact same number of rounds trying to optimize the same underlying function. The results of this posterior model check simulation allow us to judge about the extent of which the empirical results can be reproduced by the

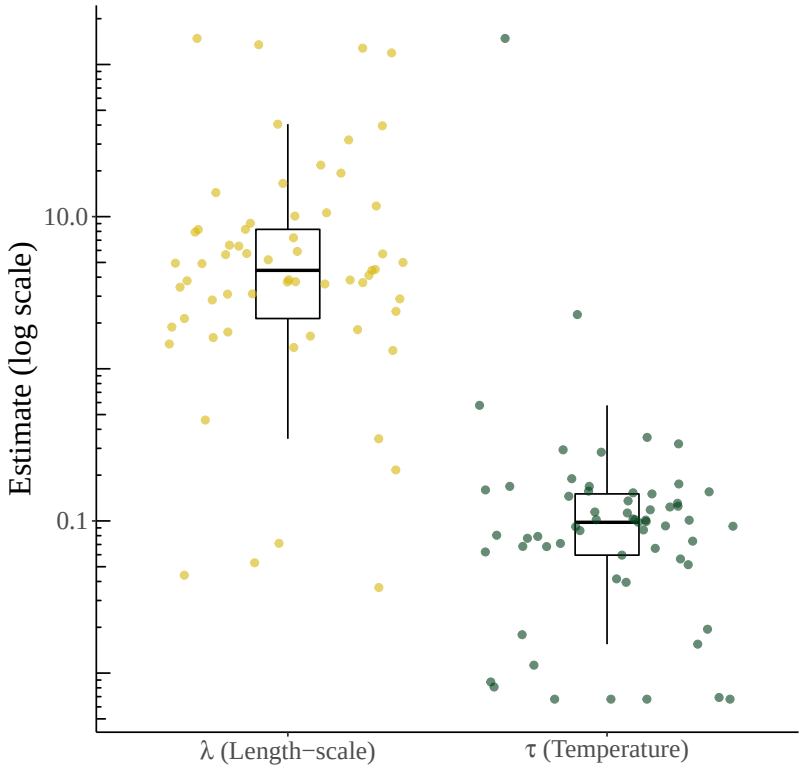

Figure 4. Parameter estimates of the Gaussian process and probability of being safe (GP-POS) model in Experiment 1.

GP-POS model. The results of this simulation are shown in Figure 5.

Even though the GP-POS model produces a similar trajectory of mean rewards as participants, its average score per trial is somewhat below that of participants. Looking at the distribution of average trial length per round, we can see that while the two distributions are very similar overall, participants managed to more often play rounds to the maximum number of trials than the simulations. While this could indicate that participants were even more risk averse than the best fitting GP-POS models, the finding that they also outperformed the GP-POS model in score obtained shows that this did not negatively affect their performance.

Discussion

Within a first experiment assessing behavior in a univariate variant of the risky spatially correlated multi-armed bandit, we found that participants managed to learn within this tasks, improved their scores over trials, and performed better than what

would be expected by chance. Moreover, participants tended to select input points which were classified as safe or maximizing points by a rational safe optimization algorithm. Using leave-one-block-out cross-validation, we found that participants' behavior was best predicted by a probability of staying safe (POS) decision strategy that mainly cares about sampling points above the provided threshold. Overall, participants did not generalize far beyond the already observed input points, resulting in roughly equal performance of both the option learning and the Gaussian process regression model. There seemed relatively little effect of the level of riskiness in the task (manipulated by the level of the safe threshold). Perhaps participants start behaving equally risk averse once a threshold is introduced. Overall, participants performed slightly better and chose inputs in a more risk averse manner than predicted by the GP-POS model (parameterized by participant's estimates), although the model produced behavior very similar to participants' overall.

Experiment 2: Bivariate inputs

In the second experiment, participants were asked to maximize an unknown bivariate function, which was represented by a two-dimensional grid world (Fig. 6). Moreover, we introduced a standard, risk-free condition as a within-subjects factor to see if participants can switch between the two different modes or riskiness.

Participants

62 participants (37 male), with an average age of 31.77 years (SD = 8.97) were recruited via Amazon Mechanical Turk and received $1 for their participation and a performance-dependent bonus of up to $1. The average completion time of the whole experiment was 11 minutes.

Procedure

We created functions $f\colon \mathbf{x} \mapsto y$ with $\mathbf{x} = (x_1, x_2)^\top$, defined over the grid $x_1, x_2 \in [0, 0.1, \ldots, 1]$ resulting in a 11×11 grid, with $y = f(\mathbf{x}) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$. As in Experiment 1, on each round the function f was sampled independently from a GP

with an RBF kernel (length-scale $\lambda = 2$). The output values y varied between 0 and 100 and one initial input point with an output above the threshold of 50 was chosen at random and provided at the start of each round. We varied the level of risk within-participants: out of the total of 10 rounds there were 5 which were "safe" (i.e., unconstrained maximization tasks without a threshold) and 5 which were labeled as "risky" (i.e., constrained maximization tasks where obtaining an output below 50 caused the round to terminate immediately, forfeiting any remaining trials in the round). The rounds were presented in a counter balanced order.

Participants were paid a basic fee of $1 and an additional bonus of $0.01 for every additional 10 points they earned overall.

Behavioral results

Figure 7 shows the results of Experiment 2. Participants performed better than chance overall ($t(61) = 15.48$, $p < .001$, $d = 1.97$). On average, participants did not increase their scores significantly over rounds (mean correlation: $r = -.04$, $t(61) = -1.34$, $p > .1$, $d = 0.16$). However, looking separately at the riskiness conditions showed that while this was true for the safe conditions (mean correlation: $r = .02$, $t(61) = 0.99$, $p > .3$, $d = 0.13$), participants did get significantly better over trials within the risky condition (mean correlation: $r = .10$, $t(61) = 2.23$, $p < .05$, $d = 0.28$). Surprisingly, participants actually scored higher in risky rounds as compared to safe rounds ($t(61) = 9.78$, $p < .001$, $d = 1.24$). This seems to be driven by a tendency towards greater exploration in the safe rounds which also explains why the average of their first sampled output considerably drops from the average revealed value (see Figure 7). In risky rounds, participants avoided scoring below the threshold for longer than was expected by chance ($t(61) = 8.06$, $p < .001$, $d = 1.02$).

Participants again explored more locally than expected by chance ($t(61) = 18.43$, $p < 0.001$, $d = 2.34$), but did not explore more locally during risky as compared to safe rounds ($t(61) = -0.31$, $p > 0.7$, $d = -0.06$).

Categorization of decisions

We again fitted a mixed-effects logistic regression analysis to participants choices and found that the best possible model contained a random intercept over participants as well as an interaction term between the effect of the safe sets and the current riskiness condition (Table 2). As expected, the effect of the safe set was again the strongest overall (Estimate = 1.85). As before, participants also chose points that were classified as maximizers more frequently (Estimate = 1.16). Additionally, being within the expansion set also deemed points to be significantly more frequently chosen this time, although this effect turned out to be relatively small (Estimate = 0.13). The interaction effect between the riskiness condition and the safe sets indicate that participants are less likely to sample from within the safe sets in the conditions without additional risks (Estimate = -0.61). This is intuitive as they are not required to focus on sampling above 50 in this condition.

Table 2
Fixed effects of the mixed-effects logistic regression in Experiment 2.

	Estimate	S.E.	z value	Pr(>|z|)
Intercept	-5.17	0.02	-271.95	0.00
Maximizer	1.16	0.11	10.58	0.00
Expander	0.13	0.06	2.24	0.02
Safe	1.85	0.11	16.75	0.00
Safe × Condition	-0.61	0.06	-10.62	0.00

Trial-by-trial modeling

Model fits for the two riskiness conditions are shown in Figure 8. We can see that the predictions of the models were generally better for the risky condition than the safe conditions ($t(61) = 9.69$, $p < .001$, $d = 1.23$).

Only analyzing the safe condition, we found that the Gaussian process regression model lead to better predictions than the option learning model ($t(61) = 4.89$, $p < .001$, $d = 0.62$). There were no differences between the different decision strategies when paired with the Gaussian process model for the safe conditions (all $p > 0.05$).

Looking at the results for the risky condition, the Gaussian process regression again predicted behavior better than the option learning model ($t(61) = 4.53$, $p < .001$, $d = 0.58$). Importantly, the probability of being safe sampling strategy lead to significantly better prediction than any other sampling strategy ($t(61) = 3.37$, $p < .01$, $d = 0.43$). Therefore, participants seem to adapt their sampling strategy to the risky constraints of the task.

Figure 9 shows the median parameter estimates for each participant for the GP-POS model(Gaussian process learning paired with the probability of being safe decision strategy) for both the risky and safe conditions. Whereas the estimates of the softmax temperature parameter differed between the two conditions ($t(61) = -4.13$ $p < .001$, $d = 0.52$) indicating more precise predictions for the risky (median estimate $\hat{\tau} = 0.11$) than for the safe condition (median estimate $\hat{\tau} = 0.27$), estimates of the length-scale λ did not differ significantly between conditions ($t(61) = -0.66$, $p = .51$, $d = 0.08$). Instead, median estimates of λ per participant correlated significantly between conditions ($r = .36$ $t(60) = 2.97$, $p < .01$), indicating that participants approached both conditions with a similar tendency towards generalization.

As in Experiment 1, we performed a posterior model check by parameterizing the GP-POS model with the participant-wise parameter estimates and let the model perform the exact same task as participants. Results of this simulation are shown in Figure 10.

We can see that, in the risky condition, the average score of the GP-POS model drops on the second trial but afterwards increases more sharply than participants' mean trajectory. From round 9 onwards, the model performs slightly better than participants. This indicates that the GP-POS model explores more extensively than participants did, incurring an initial hit to performance in order to reap later benefits. For the safe condition, the model corresponds almost perfectly with participants' mean trajectories. This is expected as this result is primarily driven by a higher temperature parameter τ, leading to an increase in random exploration. The histogram of round length again indicates that participants' behavior is more risk averse than that of the GP-POS

model, as participants managed to play for the maximum number of trials more frequently than he GP-POS model.

Discussion

Within a bivariate version of the risky spatially correlated multi-armed bandit, we found that participants improved over trials within the risky but not the safe condition, gained better scores than would be expected from random choices, played for longer than expected by chance in the risky conditions, and seemed to generalize further than in the previous task with univariate inputs. Our mixed effects regression analysis revealed that participants selected safe and maximizing points frequently, only marginally cared about choosing inputs from the expanding set input, and focused less on safe input points during safe conditions. The model comparison results showed that the Gaussian process regression model predicted behavior best in both conditions, even though predictions turned out to be generally better for the risky rounds than for the safe rounds. Importantly, whereas the different decision strategies predicted behavior equally well for the safe conditions, the probability of staying safe predicted behavior best for the risky condition. These results suggest that participants adapted their decision strategy to the task requirements while learning and generalizing about the underlying function in a similar fashion. In a posterior predictive check for the GP-POS model, we found that participants explored even more risk-aversely in the risky condition than predicted by the GP-POS model.

General discussion and conclusion

Learning unknown functions and exploiting this knowledge to maximize rewards are essential cognitive skills. Here we focused on a risky version of the spatially correlated bandit task, in which outcomes below a given threshold need to be avoided. We first analyzed participants' choices using a rational Gaussian process safe optimization strategy that establishes a safe set and tries to maximize outputs or expand the safe set by choosing inputs from this set. We found that participants shunned risks by focusing on maximizing outputs locally to "tried-and-tested" inputs,

mostly ignoring lesser known input points which could expand the safe set. This focus on avoiding unsafe inputs is consistent with a biological homeostasis maintenance principle (Korn & Bach, 2015) that prioritizes not losing everything over gaining maximum rewards (Houston, McNamara, & Hutchinson, 1993).

The results of our cross-validation model comparison revealed that participants learn and perform generalization in a similar fashion, in scenarios with and without risky constraints. While participants seemed to learn a similar representation of the reward function (using the same learning strategy) across the different task demands in Experiment 2, they did adapt their decision strategy to the riskiness of the environment, sometimes even more than predicted by the best currently available model. This in turn suggests a flexible mechanism that can guide people through risky environments via generalization and adaptive search.

In future work, we aim to focus on the factors which drive participants to switch from pure exploration to safe search strategies, and the situations in which switching constitutes a normative strategy, for example because it minimizes costs (Bach, 2015). Another promising avenue for future research will be marrying the powerful methods of generalization put forward here with restricted methods of planning that have been studied in the human reinforcement learning (Huys et al., 2015; Solway & Botvinick, 2015). Furthermore, investigating how different clinical populations differ in their search and generalization behavior when confronted with risky decision making tasks promises to extend our notion of computational mechanisms involved in mental illnesses (Huys, Maia, & Frank, 2016; Montague, Dolan, Friston, & Dayan, 2012).

Unlike previous work on human behavior in a bandit setting, which has primarily focused on pure exploration and exploitation, our work addresses a relatively novel facet—optimizing risky functions while staying above a threshold. We expect that this new approach will continue to enrich our understanding of how people resourcefully optimize outcomes in the real world.

Acknowledgments

We thank Dominik Bach and Felix Berkenkamp for helpful comments and discussions. E.S. was supported by a postdoctoral fellowship from the Harvard Data Science Initiative.

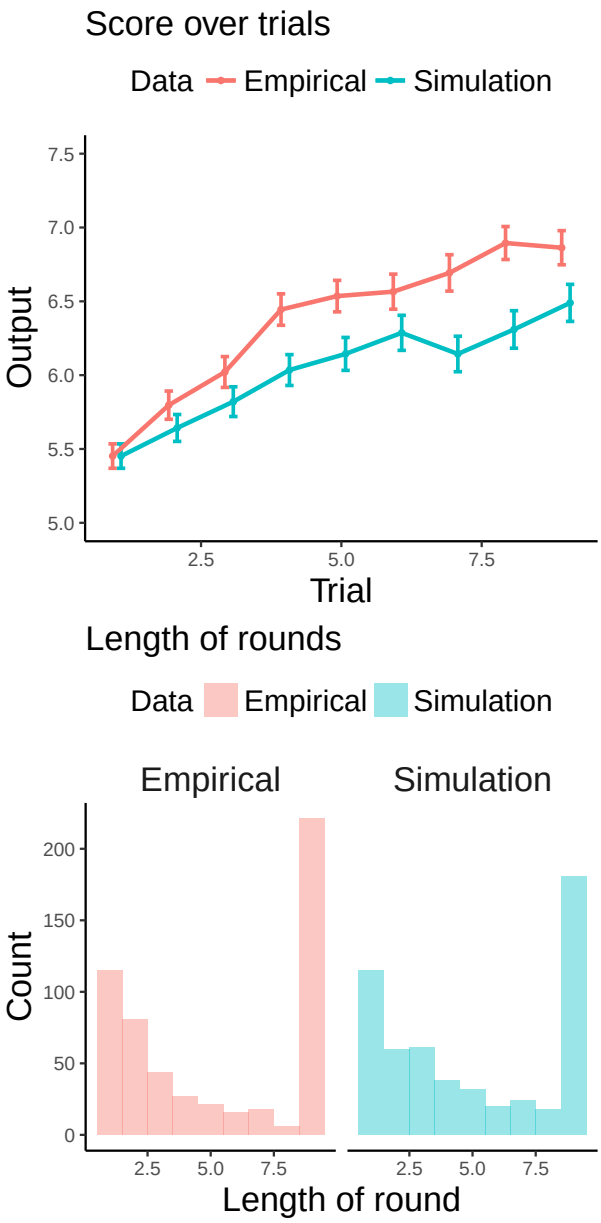

Figure 5. Posterior model checks of GP-POS model for Experiment 1. Upper panel shows the mean trajectories of scores over trials for both human participants and the GP-POS model when performing the exact same task. Lower panel shows the histogram of lengths per round, i.e. how long both participants and the model managed to sample without sampling below the threshold.

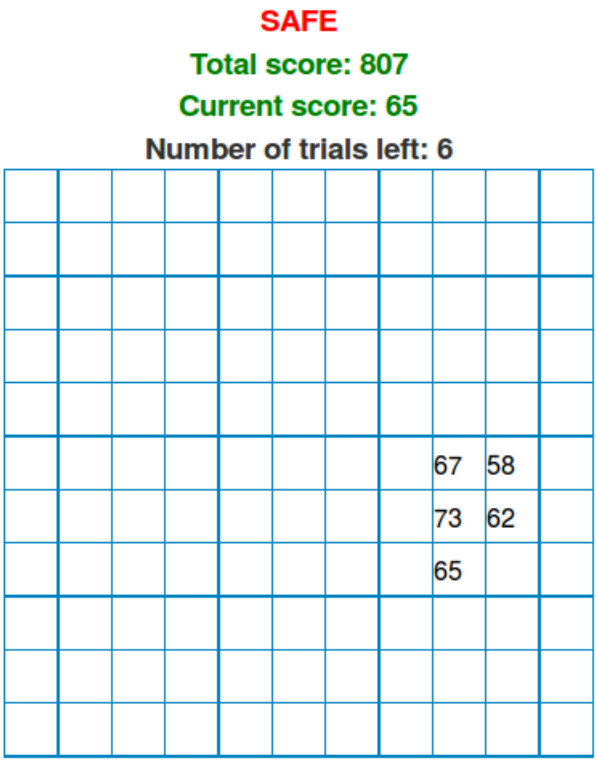

Figure 6. Screenshot of Experiment 2. Inputs were arranged in an 11 by 11 grid. Participants chose inputs by clicking on the corresponding tile, trying to choose inputs which produce high rewards. The "SAFE" condition indicates that they do not have to worry about obtaining inputs above the safe threshold of 50. Rounds at which they had to obtain outputs above 50 were marked as "RISKY".

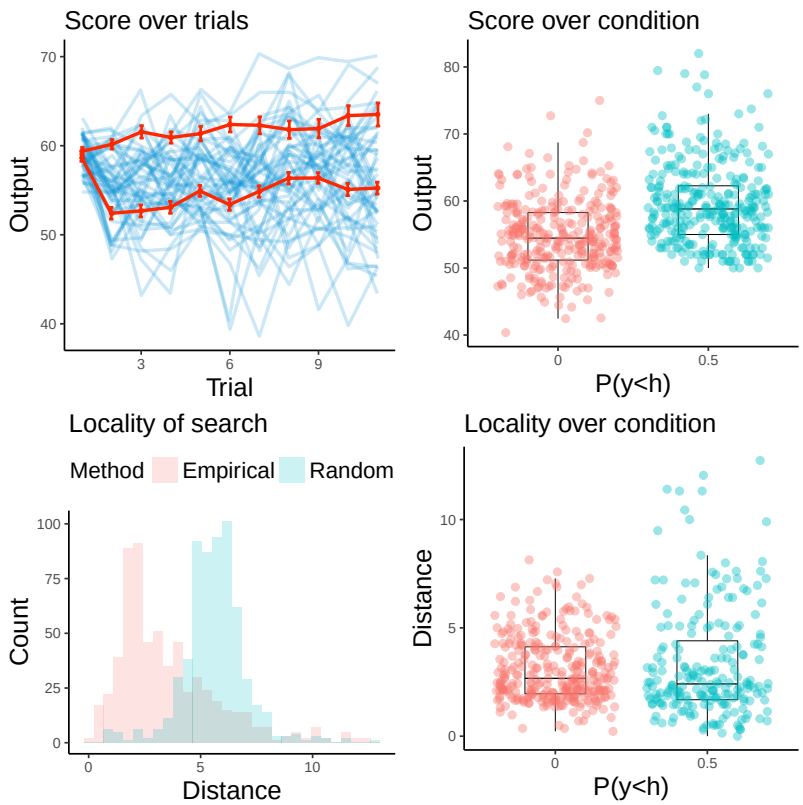

Figure 7. Results of Experiment 2. The upper left panel shows participants' scores over trials, including the average score (red lines, upper shows mean of risky condition, lower line of safe rounds) and standard errors in error bars. The upper right panel shows the average scores in safe and risky rounds. The lower left panel shows the locality of chosen inputs as compared to a random sampler. The lower right panel shows the locality of chosen inputs in safe and risky rounds.

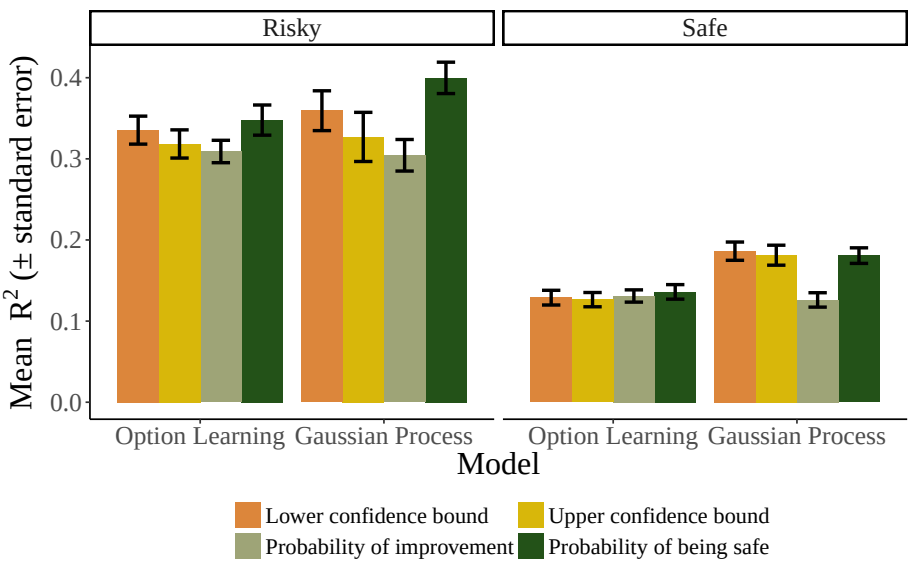

Figure 8. Trial-by-trial modeling results for Experiment 2.

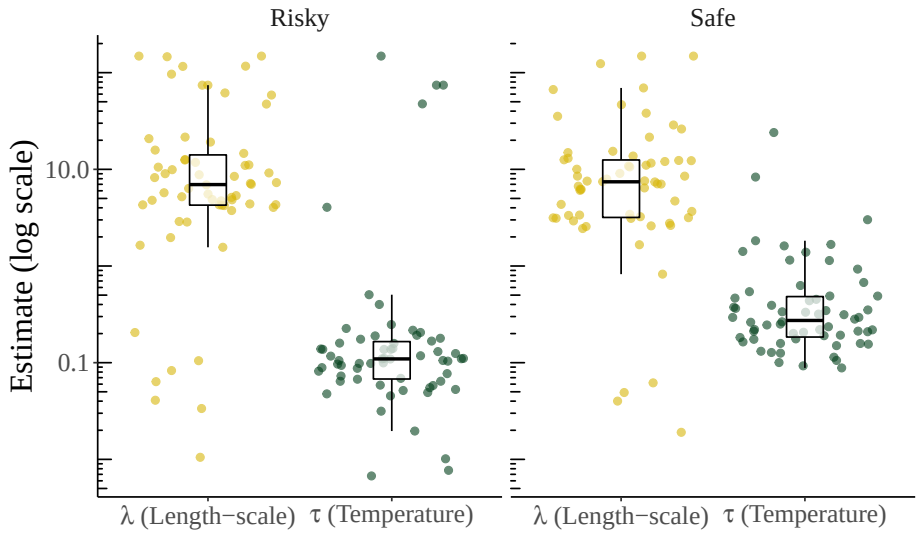

Figure 9. Parameter estimates of the GP-POS model in Experiment 2.

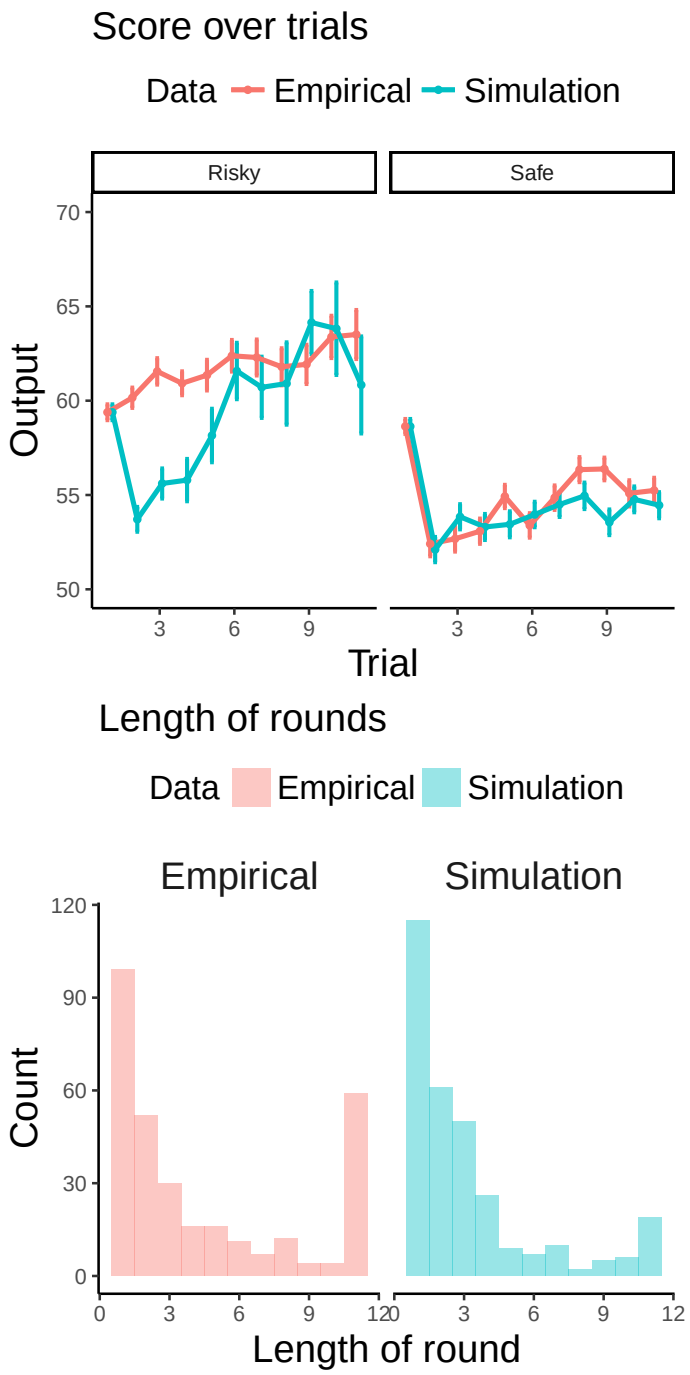

Figure 10. Posterior model checks of GP-POS model for Experiment 2. Upper panel shows the mean trajectories of scores over trials for both human participants and the GP-POS in the safe and risky rounds. Lower panel shows the histogram of lengths per round, i.e. how long both participants and the model managed to sample without sampling below the threshold.

References

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*(Nov), 397–422.

Bach, D. R. (2015). Anxiety-like behavioural inhibition is normative under environmental threat-reward correlations. *PLoS Comput Biol*, *11*(12), e1004646.

Berkenkamp, F., & Schoellig, A. P. (2015). Safe and robust learning control with gaussian processes. In *Proc. of the european control conference (ecc)* (pp. 2501–2506).

Berkenkamp, F., Turchetta, M., Schoellig, A. P., & Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. *arXiv preprint arXiv:1705.08551*.

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.

Gershman, S. J., Malmaud, J., & Tenenbaum, J. B. (2017). Structured representations of utility in combinatorial domains. *Decision*, *4*(2), 67.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 148–177.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–229.

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).

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*,

21(4), 263–268.

Houston, A. I., McNamara, J. M., & Hutchinson, J. M. (1993). General results concerning the trade-off between gaining energy and avoiding predation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *341*(1298), 375–397.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*, 3098–3103.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404.

Korn, C. W., & Bach, D. R. (2015). Maintaining homeostasis by decision-making. *PLOS Computational Biology*, *11(5):e1004301*.

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*(1), 97–106.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, *6*(1), 4–22.

Laureiro-Martínez, D., Brusoni, S., & Zollo, M. (2010). The neuroscientific foundations of the exploration- exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, *3*(2), 95.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY, USA: Henry Holt and Co., Inc.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80.

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2009). Deoptim: An r package for global optimization by differential evolution.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562.

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

Robbins, H. (1985). Some aspects of the sequential design of experiments. In *Herbert robbins selected papers* (pp. 169–177). Springer.

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., Konstantinidis, E., & Speekenbrink, M. (2017). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 081091.

Schulz, E., Speekenbrink, M., & Krause, A. (2016). A tutorial on gaussian process regression with a focus on exploration-exploitation scenarios. *bioRxiv*, 095190.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2016). Compositional inductive biases in function learning. *bioRxiv*, 091298.

Shepard, R. N., et al. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(37), 11708–11713.

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. *IEEE Transactions on Information Theory*, *58*(5), 3250–3265.

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.

Sui, Y., Gotovos, A., Burdick, J. W., & Krause, A. (2015). Safe exploration for optimization with Gaussian Processes. In *International conference on machine learning (icml).*

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017a). Exploration and generalization in vast spaces. *bioRxiv*, 171371.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017b). Mapping the unknown: The spatially correlated multi-armed bandit. In *Proceedings of the 39th annual conference of the cognitive science society.*