Exploration and generalization in vast spaces

Charley M. Wu[1,*], Eric Schulz[2], Maarten Speekenbrink[3], Jonathan D. Nelson[4,5], and Björn Meder[1,5]

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[2]Department of Psychology, Harvard University, Cambridge, Massachusetts, USA
[3]Department of Experimental Psychology, University College London, London, UK
[4]School of Psychology, University of Surrey, Guildford, UK
[5]MPRG iSearch, Max Planck Institute for Human Development, Berlin, Germany
[*]cwu@mpib-berlin.mpg.de

ABSTRACT

From foraging for food to learning complex games, many aspects of human behaviour can be framed as a search problem with a vast space of possible actions. Under finite search horizons, optimal solutions are generally unobtainable. Yet how do humans navigate vast problem spaces, which require intelligent exploration of unobserved actions? Using a variety of bandit tasks with up to 121 arms, we study how humans search for rewards under limited search horizons, where the spatial correlation of rewards (in both generated and natural environments) provides traction for generalization. Across a variety of different probabilistic and heuristic models, we find evidence that Gaussian Process function learning—combined with an optimistic Upper Confidence Bound sampling strategy—provides a robust account of how people use generalization to guide search. Our modelling results and parameter estimates are recoverable, and can be used to simulate human-like performance, providing novel insights about human behaviour in complex environments.

Introduction

Many aspects of human behaviour can be understood as search tasks[1], from foraging for food or resources[2], to searching through a hypothesis space to learn causal relationships[3], or more generally, learning which actions lead to rewarding outcomes[4]. In a natural setting, these tasks come with a vast space of possible actions, each corresponding to some reward that can only be observed through experience. In such problems, one must learn to balance the dual goals of exploring unknown options, while also exploiting familiar options for immediate returns. This frames the *exploration-exploitation dilemma*, typically studied using the multi-armed bandit framework[*5–8], with the assumption that each option has its own reward distribution to be learned independently. Yet under real-world constraints of limited time or resources, it is not enough to know *when* to explore, but also *where*.

Human learners are incredibly fast at adapting to unfamiliar environments, where the same situation is rarely encountered twice[9,10]. This highlights an intriguing gap between human and machine learning, where traditional approaches to reinforcement learning typically learn about the distribution of rewards for each state independently[4]. Such an approach falls short in more realistic scenarios where the size of the problem space is far larger than the search horizon, and it becomes infeasible to observe all possible options[11,12]. What strategies are available for an intelligent agent—biological or machine—to guide the efficient exploration of options when not all options can be explored?

[*]The multi-armed bandit is a metaphor for a row of slot machines in a casino, where each slot machine has an independent payoff distribution. Solutions to the problem propose different policies for how to learn about which arms are better to play (exploration), while also playing known high-value arms to maximize reward (exploitation).

One method for dealing with vast state spaces is to use *function learning* as a mechanism to generalize prior experience to unobserved states[13]. The function learning approach approximates a global value function over all options, including ones not experienced yet[10]. This allows for generalization to vast and potentially infinite state spaces, based on a small number of observations. Additionally, function learning scales to problems with complex sequential dynamics and has been used in tandem with restricted search methods, such as Monte Carlo sampling, for navigating intractably large search trees[14,15]. While restricted search methods have been proposed as models of human reinforcement learning in planning tasks[16,17], here we focus on situations in which a rich model of environmental structure supports learning and generalization[18].

Function learning has been successfully utilized for adaptive generalization in various machine learning applications[19,20], although relatively little is known about how humans generalize *in vivo* (e.g., in a search task; but see[8,21]). Building on previous work exploring inductive biases in pure function learning contexts[22,23] and human behaviour in univariate function optimization[24], we present the first comprehensive approach using a robust computational modelling framework to understand how humans generalize in an active search task.

Across threes studies using uni- and bivariate multi-armed bandits with up to 121 arms, we compare a diverse set of computational models in their ability to predict individual human behaviour. In all experiments, the majority of subjects are best captured by a model combining function learning using Gaussian Process (GP) regression, with an optimistic Upper Confidence Bound (UCB) sampling strategy that directly balances expectations of reward with the reduction of uncertainty. Importantly, we recover meaningful and robust estimates of the nature of human generalization, showing the limits of traditional models of associative learning[25] in tasks where the environmental structure supports learning and inference. The main contributions of this paper are threefold:

1. We introduce the *spatially correlated multi-armed bandit* paradigm as a method for studying the extent to which people use generalization to guide search, in a far more complex problem space than traditionally used to study human behaviour.

2. We find that a Bayesian model of function learning robustly captures how humans generalize and learn about the structure of the environment, where an observed tendency towards undergeneralization is shown to sometimes be beneficial.

3. We show that participants solve the exploration-exploitation dilemma by optimistically inflating expectations of reward by the underlying uncertainty, with recoverable evidence for the separate phenomena of directed (towards reducing uncertainty) and undirected (noisy) exploration.

Results

A useful inductive bias in many real world search tasks is to assume spatial correlation between rewards[26] (i.e., clumpiness of resource distribution[27]). This is equivalent to assuming that similar actions or states will yield similar outcomes. We present human data and modelling results from three experiments (Fig. 1) using univariate (Experiment 1) and bivariate (Experiment 2) environments with fixed levels of spatial correlations, and also real-world environments where spatial correlations occur naturally (Experiment 3). The spatial correlation of rewards provides a context to each arm of the bandit, which can be learned and used to generalize to yet unobserved options, thereby guiding search decisions. Additionally, since recent work has connected both spatial and conceptual representations to a common neural substrate[28],

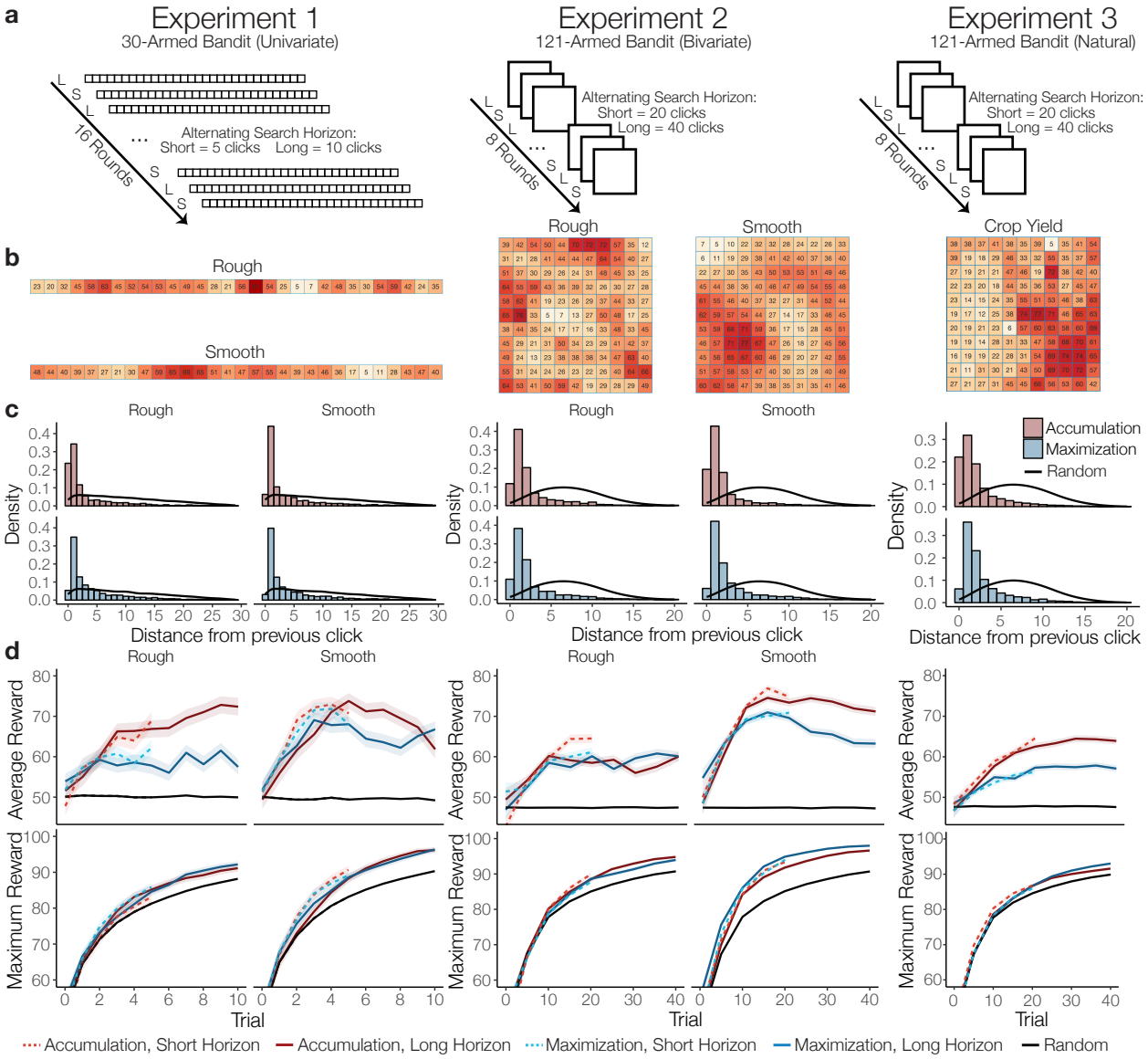

Figure 1. Procedure and behavioural results. Experiments 1 and 2 used a 2×2 between-subject design, manipulating the type of environment (Rough or Smooth) and the payoff condition (Accumulation or Maximization), while Experiment 3 manipulated only payoff conditions (between subjects) and used a set of natural environments where rewards reflect normalized crop yields from various agricultural datasets. **a)** Experiment 1 used a 1D array of 30 possible options, while Experiments 2 and 3 used a 2D array (11×11) with 121 options. Experiments took place over 16 (Exp. 1) or 8 rounds (Exp. 2 and 3), with a new environment sampled without replacement for each round. Search horizons alternated between rounds, with the horizon order counter-balanced between subjects. **b)** Examples of fully revealed search environments, where tiles were initially blank at the beginning of each round, except for a single randomly revealed tile. Rough and Smooth environments differed in the extent of spatial correlations, while Crop Yield environments have no fixed level of correlation (see SI). **c)** Locality of sampling behaviour compared to a random baseline simulated over 10,000 rounds (black line), where distance is measured using Manhattan distance. **d)** Average reward earned (Accumulation goal) and maximum reward revealed (Maximization goal), where coloured lines indicate the assigned payoff condition and shaded regions show the standard error of the mean. Short horizon trials are indicated by lighter colours and dashed lines, while black lines are a random baseline simulated over 10,000 rounds.

our results in a spatial domain provide potential pathways to other search domains, such as contextual[29–31] or semantic search[32,33].

Experiment 1

Participants ($n = 81$) searched for rewards on a 1×30 grid world, where each tile represented a reward-generating arm of the bandit (Fig. 1a). The mean rewards of each tile were spatially correlated, with stronger correlations in *Smooth* than in *Rough* environments (between subjects; Fig. 1b). Participants were either assigned the goal of accumulating the largest average reward (*Accumulation condition*), thereby balancing exploration-exploitation, or of finding the best overall tile (*Maximization condition*), an exploration goal directed towards finding the global maximum. Additionally, the search horizons alternated between rounds (*Short* = 5 vs. *Long* = 10), with the order counter-balanced between subjects. We hypothesized that if function learning guides search behaviour, participants would perform better and learn faster in smooth environments, in which stronger spatial correlations reveal more information about nearby tiles[34].

Looking first at sampling behaviour, the distance between sequential choices was more localized than chance ($t(80) = -39.8$, $p < .001$, $d = -4.4$; Fig. 1c)[†], as has also been observed in semantic search[32] and causal learning[3] domains. Participants in the Accumulation condition sampled more locally than those in the Maximization condition ($t(79) = -3.33$, $p = .001$, $d = -0.75$), corresponding to the increased demand to exploit known or near-known rewards. Comparing performance in different environments, the learning curves in Fig. 1d show that participants in Smooth environments obtained higher average rewards than participants in Rough environments ($t(79) = 3.58$, $p < .001$, $d = 0.8$), consistent with the hypothesis that spatial patterns in the environment can be learned and used to guide search. Surprisingly, longer search horizons (solid vs. dashed lines) did not lead to higher average reward ($t(80) = 0.60$, $p = .549$, $d = 0.07$). We analyzed both average reward and the maximum reward obtained for each subject, irrespective of their payoff condition (Maximization or Accumulation). Remarkably, participants in the Accumulation condition performed best according to both performance measures, achieving higher average rewards than those in the Maximization condition ($t(79) = 2.89$, $p = .005$, $d = 0.65$), and performing equally well in terms of finding the largest overall reward ($t(79) = -0.73$, $p = .467$, $d = -0.16$). Thus, a strategy balancing exploration and exploitation—at least for human learners—may achieve the global optimization goal *en passant*.

Experiment 2

Experiment 2 had the same design as Experiment 1, but used a 11×11 grid representing an underlying bivariate reward function (Fig. 1 centre) and longer search horizons to match the larger search space (*Short* = 20 vs. *Long* = 40). We replicated the main results of Experiment 1, showing that participants ($n = 80$) sampled more locally than a random baseline ($t(79) = -50.1$, $p < .001$, $d = -5.6$; Fig. 1c), participants in the Accumulation condition sampled more locally than those in the Maximization condition $t(78) = -2.75$, $p = .007$, $d = -0.61$), and overall, participants obtained higher rewards in Smooth environments than in Rough environments ($t(78) = 6.55$, $p < .001$, $d = 1.47$; Fig. 1d). For both locality of sampling and the difference in average reward between environments, the effect size was larger in Experiment 2 than in Experiment 1. We also replicated the result that participants in the Accumulation condition were as good as participants in the Maximization condition at discovering the highest rewards ($t(78) = -0.62$, $p = .534$, $d = -0.14$), yet in Experiment 2 the Accumulation condition did not lead to substantially better performance than the Maximization condition in terms of average reward ($t(78) = -1.31$, $p = .192$, $d = -0.29$). Again, short search horizons led to the same level of

[†] All reported t-tests are two-sided.

performance as longer horizons, ($t(79) = -0.96$, $p = .341$, $d = -0.11$), suggesting that learning occurs rapidly and peaks rather early.

Experiment 3

Experiment 3 used the same 121-armed bivariate bandit as Experiment 2, but rather than generating environments with fixed levels of spatial correlations, we sampled environments from 20 different agricultural datasets[35], where payoffs correspond to the normalized yield of various crops (e.g., wheat, corn, and barley; see SI for details). These datasets are naturally segmented into a grid based on the rows and columns of a field, thus requiring no interpolation or other transformation of the data except for the normalization of payoffs. The crucial difference from Experiment 2 is that there is no fixed level of spatial correlation across the entire search task, although the assumption of spatial correlations is still a reasonable heuristic for guiding search.

As in both previous experiments, participants ($n = 80$) sampled more locally than random chance ($t(79) = -50.1$, $p < .001$, $d = -5.6$), with participants in the Accumulation condition sampling more locally than those in the Maximization condition ($t(78) = -3.1$, $p = .003$, $d = -0.69$). In the natural environments, we found accumulation participants achieved a higher average reward than maximization participants ($t(78) = 2.7$, $p = .008$, $d = 0.61$), with an effect size similar to Experiment 1. There was no difference in maximum reward across payoff conditions ($t(78) = 0.3$, $p = .8$, $d = 0.06$) as in all previous experiments, showing that the goal of balancing exploration-exploitation dominates by both performance metrics. In contrast to the previous experiments, we found that a longer search horizon lead to slightly higher average rewards ($t(78) = 2.1$, $p = .04$, $d = 0.21$), although taken together with the previous experiments, performance under short horizons proved highly efficient.

Modelling Generalization and Search

To better understand how participants explored the environment, we compared a diverse set of computational models in their ability to predict each subject's trial-by-trial choices (see Fig. S1 and Table S2 for full results). These models include different combinations of *models of learning* and *sampling strategies*, which map onto the distinction between belief and sampling models, central to theories in statistics[36], psychology[37], and philosophy of science[38]. Models of learning form inductive beliefs about the value of possible options (including unobserved options) based on previous observations, while sampling strategies transform these beliefs into probabilistic predictions about where a participant will sample next. We also consider *simple heuristics*, which are competitive models of human behaviour in bandit tasks[5] yet do not maintain a model of the world (see SI). By far the best predictive models used *Gaussian Process* (GP) regression[39,40] as a mechanism for generalization, and *Upper Confidence Bound* (UCB) sampling[41] as an optimistic solution to the exploration-exploitation dilemma.

Function learning provides a possible explanation of how individuals generalize from previous experience to unobserved options, by adaptively learning an underlying function mapping options onto rewards. We use GP regression as an expressive model of human function learning, which has known equivalencies to neural network function approximators[42], yet provides psychologically interpretable parameter estimates about the extent to which generalization occurs. GP function learning can guide search by making predictions about the expected mean $m(\mathbf{x})$ and the associated uncertainty $s(\mathbf{x})$ (estimated here as a standard deviation) for each option $\mathbf{x}$ in the global state space (see Fig. 2b), conditioned on a finite number of previous observations of rewards $\mathbf{y}_T = [y_1, y_2, \ldots, y_T]^\top$ at inputs $\mathbf{X}_T = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$. Similarities between options are modelled by a *Radial Basis Function* (RBF) kernel:

$$k_{\text{RBF}}(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{\lambda}\right), \tag{1}$$

where λ governs how quickly correlations between points $\mathbf{x}$ and $\mathbf{x}'$ (e.g., two tiles on the grid) decay towards zero as their distance increases. We use λ as a free parameter, which can be interpreted psychologically as the extent to which people generalize spatially. Since the GP prior is completely defined by the RBF kernel, the underlying mechanisms are similar to Shepard's universal gradient of generalization[43], which also models generalization as an exponentially decreasing function of distance between stimuli. To illustrate, generalization to the extent of $\lambda = 1$ corresponds to the assumption that the rewards of two neighbouring options are correlated by $r = 0.61$, and that this correlation decays to zero if options are further than three tiles away from each other. Smaller λ values would lead to a more rapid decay of assumed correlations as a function of distance.

Given estimates about expected rewards $m(\mathbf{x})$ and the underlying uncertainty $s(\mathbf{x})$ from the function learning model, UCB sampling produces valuations of each option $\mathbf{x}$ using a simple weighted sum:

$$\text{UCB}(\mathbf{x}) = m(\mathbf{x}) + \beta s(\mathbf{x}), \tag{2}$$

where β is a free parameter governing how much the reduction of uncertainty is valued relative to expectations of reward (Fig. 2c). To illustrate, an exploration bonus of $\beta = 0.5$ suggests participants would prefer the hypothetical option x_1 predicted to have mean reward $m(x_1) = 60$ and standard deviation $s(x_1) = 10$, over option x_2 predicted to have mean reward $m(x_2) = 64$ and standard deviation $s(x_2) = 1$. This is because sampling x_1 is expected to reduce a large amount of uncertainty, even though x_2 has a higher mean reward ($\text{UCB}(x_1) = 65$ vs. $\text{UCB}(x_2) = 64.5$). This trade-off between exploiting known high-value rewards and exploring to reduce uncertainty[44] can be interpreted as optimistically inflating expectations of reward by the attached uncertainty, and can be contrasted to two separate sampling strategies that only sample points based on high expected reward (Pure Exploitation) or high uncertainty (Pure Exploration).

$$\text{PureExploit}(\mathbf{x}) = m(\mathbf{x}) \tag{3}$$

$$\text{PureExplore}(\mathbf{x}) = s(\mathbf{x}) \tag{4}$$

Figure 2 shows how the GP-UCB model makes inferences about the search space, and uses UCB sampling (combined with a softmax choice rule) to make probabilistic predictions about where the participant will sample next. We refer to this model as the *Function Learning Model* and contrast it with an *Option Learning Model*. The Option Learning Model uses a Bayesian mean tracker (BMT) to learn about the distribution of rewards for each option independently (see Methods). The Option learning Model is a traditional associative learning model, and can be understood as a variant of a Kalman filter where rewards are assumed to be time-invariant[6]. Like the Function Learning Model, the Option Learning Model also generates normally distributed predictions with mean $m(\mathbf{x})$ and standard deviation $s(\mathbf{x})$, which we combine with the same set of sampling strategies and the same softmax choice rule to make probabilistic predictions about search. For both models, we use the softmax temperature parameter (τ) to estimate the amount of undirected exploration (i.e., higher temperatures correspond to more noisy sampling), in contrast to the β parameter of UCB, which estimates the level of exploration directed towards reducing uncertainty.

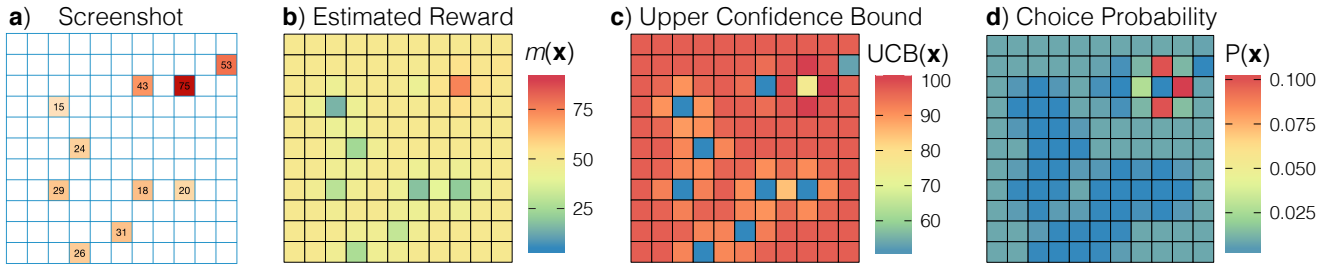

Figure 2. Overview of the Function Learning-UCB Model specified using median participant parameter estimates (see Table S2). **a**) Screenshot of Experiment 2. Participants were allowed to select any tile until the search horizon was exhausted. **b**) Estimated reward (not shown, the estimated uncertainty) as predicted by the GP Function Learning model, based on the points sampled in Panel a. **c**) Upper confidence bound of predicted rewards. **d**) Choice probabilities after a softmax choice rule. $P(\mathbf{x}) = \exp(\mathrm{UCB}(\mathbf{x})/\tau)/\sum_{j=1}^{N}\exp(\mathrm{UCB}(\mathbf{x}_j)/\tau)$, where τ is the temperature parameter (i.e., higher temperature values lead to more random sampling).

Modelling results

Experiment 1

Participants were better described by the Function Learning Model than the Option Learning Model ($t(80) = 14.01$, $p < .001$ $d = 1.56$; comparing cross-validated predictive accuracies, both using UCB sampling), providing evidence that participants generalized instead of learning rewards for each option independently. Furthermore, by decomposing the UCB sampling algorithm into Pure Exploit or Pure Explore components, we show that both expectations of reward and estimates of uncertainty are necessary components for the Function Learning Model to predict human search behaviour, with Pure Exploitation ($t(80) = -8.85$, $p < .001$, $d = -0.98$) and Pure Exploration ($t(80) = -16.63$, $p < .001$, $d = -1.85$) variants making worse predictions than the combined UCB algorithm.

Because of the observed tendency to sample locally, we created a localized variant of both Option Learning and Function Learning Models (indicated by an asterisk *; Fig. 3a), penalizing options farther away from the previous selected option (without introducing additional free parameters; see Methods). While the Option Learning* Model was better than the standard Option Learning Model ($t(80) = 16.13$, $p < .001$, $d = 1.79$), the standard Function Learning Model still outperformed its localized variant ($t(80) = -5.05$, $p < .001$, $d = 0.56$). Overall, 56 out of 81 participants were best described by the Function Learning Model, with an additional 10 participants best described by the Function Learning* Model with localization.

Figure 3b shows simulated learning curves of each model in comparison to human performance, where models were specified using parameters sampled from participant estimates (10,000 samples with replacement). Whereas both versions of the the Option Learning-UCB Model barely beat the performance of a completely random baseline, both standard and localized versions of the Function Learning-UCB Model behave sensibly and improve performance over time. This suggests there is an overlap between elements of human intelligence responsible for successful performance in our task, and elements of participant behaviour captured by the Function Learning Model.

Looking more closely at the parameter estimates of the Function Learning Model (Fig. 3c), we find that people tend to underestimate the extent of spatial correlations, with median per-participant λ estimates significantly lower than the ground truth ($\lambda_{Smooth} = 2$ and $\lambda_{Rough} = 1$) for both Smooth (Wilcoxon signed rank test; $\hat{\lambda}_{Smooth} = 0.5$, $Z = -7.1$, $p < .001$, $r = 1.1$) and Rough environments ($\hat{\lambda}_{Rough} = 0.5$, $Z = -3.4$, $p < .001$, $r = 0.55$). This can be interpreted as a tendency towards undergeneralization. Additionally, we found that the estimated exploration bonus of UCB sampling (β) was reliably greater than zero ($\hat{\beta} = 0.51$,

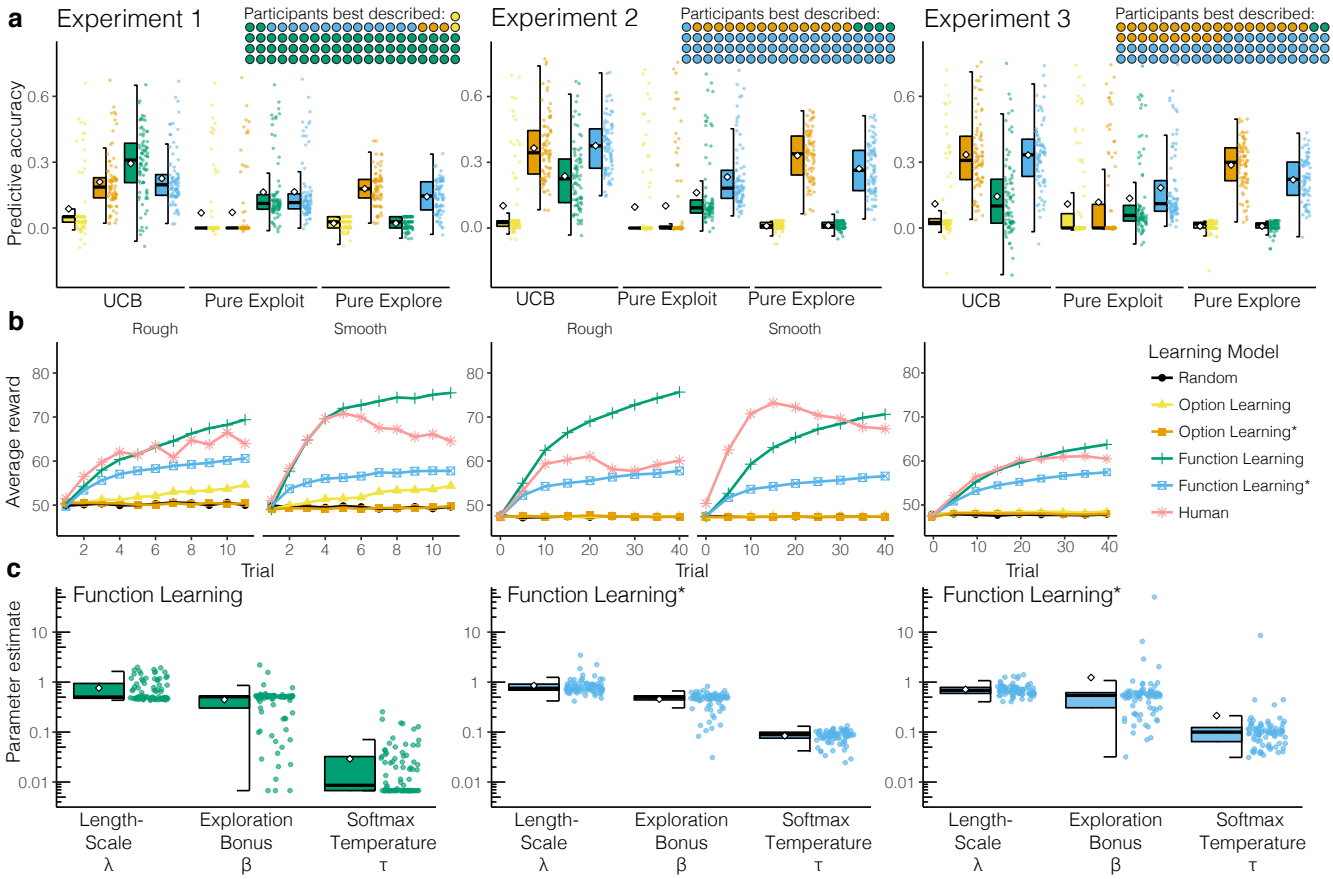

Figure 3. Modelling results. **a**) Cross-validated predictive accuracy of each model, with box plots indicating the interquartile range, the median (horizontal line), mean (diamond), and 1.5x IQR (whiskers). Each individual participant is shown as a single dot, with the number of participants best described shown as an icon array (inset; aggregated by sampling strategies). Asterisks (*) indicate a localized variant of the Option Learning or Function Learning models, where predictions are weighted by the inverse distance from the previous choice (see Methods). **b**) Averaged learning curves of participants and models (combined with UCB-sampling only) simulated over 10,000 replications using sampled per-participant parameter estimates. Learning curves (and parameter estimates) are separated by environment, aggregated over payoff conditions and search horizons. **c**) Parameter estimates of the best predicting model for each experiment. Each coloured dot is the median estimate per participant, with box plots indicating 1.5x IQR (whiskers), median (horizontal line), and mean (diamond).

$Z = -7.7$, $p < .001$, $r = 0.86$; compared to lower estimation bound), reflecting the valuation of sampling uncertain options, together with exploiting high expectations of reward. Lastly, we found relatively low estimates of the softmax temperature parameter ($\hat{\tau} = 0.01$), suggesting that the search behaviour of participants corresponded closely to selecting the very best option, once they had taken into account both the exploitation and exploration components of the available actions.

Experiment 2

In a more complex bivariate environment (Fig. 3a), the Function Learning Model again made better predictions than the Option Learning Model ($t(79) = 9.99$, $p < .001$, $d = 1.12$), which was also the case when comparing localized Function Learning* to localized Option Learning* ($t(79) = 2.05$, $p = .044$, $d = 0.23$). In the two-dimensional search environment of Experiment 2, adding localization improved predictions for both Option Learning ($t(79) = 19.92$, $p < .001$, $d = 2.23$) and Function Learning ($t(79) = 10.47$, $p < .001$, $d = 1.17$), in line with the stronger tendency towards localized sampling compared to

Experiment 1 (see Fig. 1c). Altogether, 61 out of 80 participants were best predicted by the localized Function Learning* model, whereas only 12 participants were best described by the localized Option Learning* model. Again, both components of the UCB strategy were necessary to predict choices, with Pure Exploit ($t(79) = -6.44$, $p < .001$, $d = -0.72$) and Pure Explore($t(79) = -14.32$, $p < .001$, $d = -1.60$) making worse predictions.

As in Experiment 1, the simulated learning curves of the Option Learning models performed poorly and were indistinguishable from a random sampling strategy (Fig. 3b), whereas both variants of the Function Learning Model achieved performance comparable to that of human participants. Median per-participant parameter estimates (Fig. 3c) from the Function Learning*-UCB Model showed that while participants generalized somewhat more than in Experiment 1 ($\hat{\lambda} = 0.75$, $Z = -3.7$, $p < .001$, $r = 0.29$), they again underestimated the strength of the underlying spatial correlation in both Smooth ($\hat{\lambda}_{Smooth} = 0.78$, $Z = -5.8$, $p < .001$, $r = 0.88$; comparison to $\lambda_{Smooth} = 2$) and Rough environments ($\hat{\lambda}_{Rough} = 0.75$, $Z = -4.7$, $p < .001$, $r = 0.78$; comparison to $\lambda_{Rough} = 1$). This suggests a robust tendency to undergeneralize. There were no differences in the estimated exploration bonus β between Experiment 1 and 2 ($\hat{\beta} = 0.5$, $Z = 0.86$, $p = .80$, $r = 0.07$), although the estimated softmax temperature parameter τ was larger than in Experiment 1 ($\hat{\tau} = 0.09$; $Z = -8.89$, $p < .001$, $r = 0.70$). Experiment 2 therefore replicated the main findings of Experiment 1, which when taken together, provide strong evidence that human search behaviour is best explained by a combination of function learning paired with an optimistic trade-off between exploration and exploitation.

Experiment 3

Using natural environments without a fixed level of spatial correlations, we still replicated the results that Function Learning made better predictions than Option Learning ($t(79) = 3.03$, $p = .003$, $d = 1.12$), adding localization improved predictions for both Option Learning ($t(79) = 18.83$, $p < .001$, $d = 1.24$) and Function Learning ($t(79) = 14.61$, $p < .001$, $d = 1.12$), and that the combined UCB algorithm formed better than using only a Pure Exploit ($t(79) = 12.97$, $p < .001$, $d = 0.92$) or a Pure Explore strategy ($t(79) = 5.87$, $p < .001$, $d = 0.81$). However, the difference between the localized Function Learning* and the localized Option Learning* was negligible ($t(79) = 0.32$, $p = .75$, $d = 0.01$). This is perhaps due to the high variability across environments, which makes it harder to predict out-of-sample choices using generalization behaviour (i.e., λ) estimated from a separate set of environments. Nevertheless, Function Learning* was still the best predicting model for the majority of participants (48 out of 80 participants).

Looking at the simulated learning curves in Figure 3b, we see that as in all previous experiments, both variants of Function Learning perform similar to humans, whereas the Option Learning Model produces performance no better than a random baseline, indicating that it is not a plausible model of human behaviour[45]. The parameter estimates from the Function Learning* Model are largely consistent with the results from Experiment 2 (Fig. 3c), but with participants generalizing slightly less ($\hat{\lambda}_{natural} = 0.68$, $Z = -3.4$, $p < .001$, $r = 0.27$), and exploring slightly more, with a small increase in both directed exploration ($\hat{\beta}_{natural} = 0.54$, $Z = -2.3$, $p = .01$, $r = 0.18$) and undirected exploration ($\hat{\tau}_{natural} = 0.1$, $Z = -2.2$, $p = .02$, $r = 0.17$) parameters. Altogether, the parameter estimates are highly similar to the previous experiments.

Robustness and Recovery

We conducted both model and parameter recovery simulations to assess the validity of our modelling results (see SI for full details). We performed model recovery by simulating data using a *generating model* specified by participant parameter estimates. We then used the same cross-validation procedure to fit a *recovering model*. In all cases, the best predictive accuracy occurred when the recovering model matched

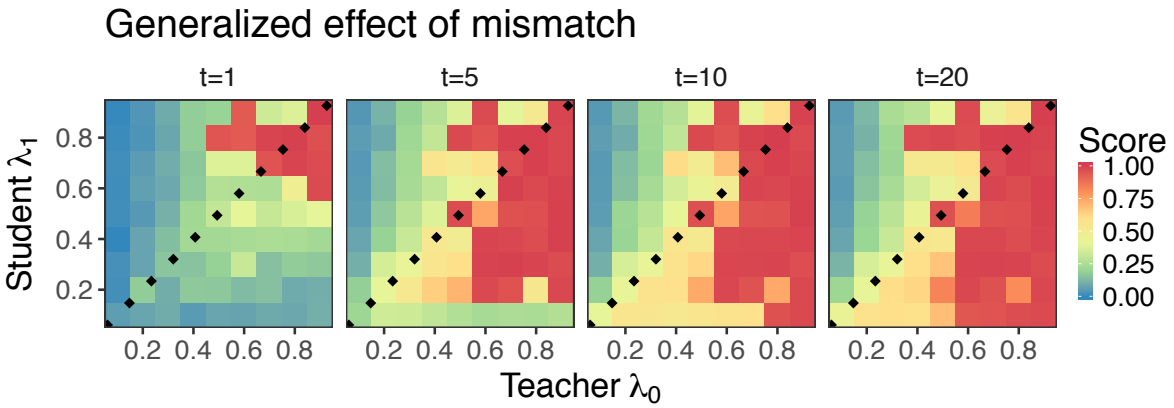

Figure 4. Mismatched length-scale (λ) simulation results. The teacher length-scale λ_0 is on the x-axis, the student length-scale λ_1 is on the y-axis, and each panel is performance at a different trial t. The teacher λ_0 values were used to generate environments, while the student λ_1 values were used to parameterize the Function Learning Model to simulate search performance. The dotted lines show where $\lambda_0 = \lambda_1$ and mark the difference between undergeneralization and overgeneralization, with points below the line indicating undergeneralization. We report the median score (over 100 replications) as a standardized measure of performance, such that 0 shows the lowest possible and 1 the highest possible log unit-performance.

the generating model (Fig. S2), suggesting robustness to Type I errors and ruling out model overfitting (i.e., the Function Learning Model did not best predict data generated by the Option Learning Model). Parameter recovery was performed to ensure that each parameter in the Function Learning-UCB Model robustly captured separate and distinct phenomena. In all cases, the generating and recovered parameter estimates were highly correlated (Fig. S3). It is noteworthy that we found distinct and recoverable estimates for β (exploration bonus) and τ (softmax temperature), supporting the existence of exploration *directed* towards reducing uncertainty[12] as a separate phenomena from noisy, undirected exploration[46].

The Adaptive Nature of Undergeneralization

In Experiments 1 and 2, we observed a robust tendency to undergeneralize compared to the true level of spatial correlations in the environment. Therefore, we ran simulations to assess how different levels of generalization influence search performance when paired with different types of environments. We found that undergeneralization largely leads to better performance than overgeneralization—and remarkably—is sometimes even better than an exact match to the underlying structure of the environment (Fig. 4). To run these simulations, we first generated search environments by sampling from a GP prior specified using a *teacher* length-scale (λ_0), and then simulating search in this environment by specifying the Function Learning-UCB Model with a *student* length-scale (λ_1). Instead of a discrete grid, we chose a set-up common in Bayesian optimization[47] with continuous bivariate inputs in the range $x, y = [0, 1]$, allowing for a broader set of potential mismatched alignments (see Fig. S4 for simulations using the exact design of each experiment).

We find that undergeneralization largely leads to better performance than overgeneralization, and that this effect is more pronounced over time t (i.e., longer search horizons). Estimating the best possible alignment between λ_0 and λ_1 revealed that underestimating λ_0 by an average of about 0.21 produces the best scores over all scenarios. These simulation results show that the systematically lower estimates of λ captured by our models is not necessarily a flaw in human cognition, but can sometimes lead to better performance. Indeed, simulations based on the natural environments used in Experiment 3 (which had no

fixed level of spatial correlations) revealed that the range of participant λ estimates were highly adaptive to the environments they encountered (Fig. S4c). Undergeneralization might not be a bug, but rather an important feature of human behaviour.

Discussion

How do people learn and adaptively make good decisions when the number of possible actions is vast and not all possibilities can be explored? We found that Function Learning, operationalized using GP regression, provides a mechanism for generalization, which can be used to guide search towards unexplored, yet promising options. Combined with Upper Confidence Bound (UCB) sampling, this model navigates the exploration-exploitation dilemma by optimistically inflating expectations of reward by the estimated uncertainty.

While GP function learning combined with a UCB sampling algorithm has been successfully applied to search problems in ecology[48], robotics[49,50], and biology[51], there has been little psychological research on how humans learn and search in environments with a vast set of possible actions. The question of how generalization operates in an active learning context is of great interest to the research community, and our work makes important theoretical and empirical contributions. Expanding on previous studies that noticed an overlap between GP-UCB and human learning rates[8,24], we present the first study to use cognitive modelling to understand how humans generalize and address the exploration-exploitation dilemma in a complex search task with spatially correlated outcomes. Through multiple analyses, including trial-by-trial predictive cross-validation and simulated behaviour using participants' parameter estimates, we competitively assessed which models best predicted human behaviour. The vast majority of participants were best described by the Function Learning-UCB model or its localized variant. Parameter estimates from the best-fitting Function Learning-UCB models suggest there was a systematic tendency to undergeneralize the extent of spatial correlations, which we found can sometimes lead to better search performance than even an exact match with the underlying structure of the environment (Fig. 4).

Altogether, our modelling framework yielded highly robust and recoverable results (Fig. S2) and parameter estimates (Fig. S3). Whereas previous research on exploration bonuses has had mixed results[6,12,46], we found recoverable parameter estimates for the separate phenomena of directed exploration encoded in UCB exploration parameter β and the noisy, undirected exploration encoded in the softmax temperature parameter τ. Even though UCB sampling is both *optimistic* (always treating uncertainty as positive) and *myopic* (only planning the next timestep), similar algorithms have competitive performance guarantees in a bandit setting[52]. This shows a remarkable concurrence between intuitive human strategies and state of the art in machine learning research.

Limitations and extensions

One potential limitation was that we failed to find systematic differences in model parameters across the two payoff conditions. The goal of balancing exploration-exploitation (Accumulation condition) or the goal of global optimization (Maximization condition) was induced through the manipulation of written instructions and comprehension check questions (see Methods). While this may have been sufficient for observing behavioural differences (i.e., performance and locality), it may have been inadequate to produce reliable differences in terms of generalization or exploration (either directed or undirected), which were the (non-exhaustive) aspects of human behaviour captured by our models. Indeed, the practical difference between these two goals is even murky in the Bayesian optimization literature, which from a purely computational perspective often proposes abandoning the strict goal of finding the global optimum[53], in favour of an approximate measure of performance, such as cumulative regret[52], which more closely aligns

to our Accumulation payoff condition. Remarkably, participants assigned to the Accumulation goal also accomplished the Maximization goal for free.

In addition to providing the best model of human behaviour, the Function Learning Model also offers many opportunities for theory integration. The Option Learning Model can itself be reformulated as special case of GP regression[54]. When the length-scale of the RBF kernel approaches zero ($\lambda \to 0$), the Function Learning Model assumes state independence, as in the Option Learning Model. Thus, there may be a continuum of reinforcement learning models, ranging from the traditional assumption of state *in*dependence to the opposite extreme, of complete state *inter*-dependence. Moreover, GPs also have equivalencies to Bayesian neural networks[42], suggesting a further link to distributed function learning models[55]. Indeed, one explanation for the impressive performance of deep reinforcement learning[14] is that neural networks are specifically a powerful type of function approximator[56].

Lastly, both spatial and conceptual representations have been connected to a common neural substrate in the hippocampus[28], suggesting a potential avenue for applying the same Function Learning-UCB model for modelling human learning using contextual[29–31], semantic[32,33], or potentially even graph-based features. One hypothesis for this common role of the hippocamus is that it performs predictive coding of future state transitions[57], also known as "successor representation"[25]. In our task, where there are no restrictions on state transitions (i.e., each state is reachable from any prior state), it may be the case that the RBF kernel driving our GP Function Learning model performs the same role as the transition matrix of a successor representation model, where state transitions are learned via a random walk policy.

Conclusions

We introduced a new paradigm for studying how people use generalization to guide the active search for rewards, and found a systematic—yet sometimes beneficial—tendency to undergeneralize. Additionally, we uncovered strong evidence for the separate phenomena of directed exploration (towards reducing uncertainty) and noisy, undirected exploration. Even though our current implementation only grazes the surface of the types of complex tasks people are able to solve—and indeed could be extended in future studies using temporal dynamics or depleting resources—it is far richer in both the set-up and modelling framework than traditional multi-armed bandit problems used for studying human behaviour. Our empirical and modelling results show how function learning, combined with optimistic search strategies, may provide the foundation of adaptive behaviour in complex environments.

Methods

Participants

81 participants were recruited from Amazon Mechanical Turk for Experiment 1 (25 Female; mean $\pm$ SD age 33 $\pm$ 11), 80 for Experiment 2 (25 Female; mean $\pm$ SD age 32 $\pm$ 9), and 80 for Experiment 3 (24 Female; mean $\pm$ SD age 35 $\pm$ 10). In all experiments, participants were paid a participation fee of \$0.50 and a performance contingent bonus of up to \$1.50. Participants earned on average \$1.14 $\pm$ 0.13 and spent 8 $\pm$ 4 minutes on the task in Experiment 1, earned \$1.64 $\pm$ 0.20 and spent 8 $\pm$ 4 minutes in Experiment 2, and earned \$1.53 $\pm$ 0.15 and spent 8 $\pm$ 5 minutes in Experiment 3. Participants were only allowed to participate in one of the experiments, and were required to have a 95% HIT approval rate and 1000 previously completed HITs. The Ethics Committee of the Max Planck Institute for Human Development approved the methodology and all participants consented to participation through an online consent form at the beginning of the survey.

Design

Experiments 1 and 2 used a 2×2 between-subjects design, where participants were randomly assigned to one of two different payoff structures (*Accumulation condition* vs. *Maximization condition*) and one of two different classes of environments (*Smooth* vs. *Rough*), whereas Experiment 3 used environments from real-world agricultural datasets, and manipulated only the payoff structure (between subject). Each grid world represented a (either uni- or bivariate) function, with each observation including normally distributed noise, $\varepsilon \sim \mathcal{N}(0,1)$. The task was presented over either 16 rounds (Exp. 1) or 8 rounds (Exp. 2 and 3) on different grid worlds drawn from the same class of environments. Participants had either a short or long search horizon (Exp. 1: $[5,10]$; Exp. 2 and 3: $[20,40]$) to sample tiles on the grid, including repeat clicks. The search horizon alternated between rounds (within subject), with initial horizon length counterbalanced between subjects.

Materials and procedure

Prior to starting the task, participants observed four fully revealed example environments and had to correctly complete three comprehension questions. At the beginning of each round, one random tile was revealed and participants could click any of the tiles in the grid until the search horizon was exhausted, including re-clicking previously revealed tiles. Clicking an unrevealed tile displayed the numerical value of the reward along with a corresponding colour aid, where darker colours indicated higher point values. Per round, observations were scaled to a randomly drawn maximum value in the range of 65 to 85, so that the value of the global optima could not be easily guessed (e.g., a value of 100). Re-clicked tiles could show some variations in the observed value due to noise. For repeat clicks, the most recent observation was displayed numerically, while hovering over the tile would display the entire history of observation. The colour of the tile corresponded to the mean of all previous observations.

Payoff conditions. We compared performance under two different payoff conditions, requiring either a balance between exploration and exploitation (*Accumulation condition*) or corresponding to consistently making exploration decisions (*Maximization condition*). In each payoff condition, participants received a performance contingent bonus of up to \$1.50. *Accumulation condition* participants were given a bonus based on the average value of all clicks as a fraction of the global optima, $\frac{1}{T}\sum(\frac{y_t}{y^*})$, where y^* is the global optimum, whereas participants in the *Maximization condition* were rewarded using the ratio of the highest observed reward to the global optimum, $(\frac{\max y_t}{y^*})^4$, taken to the power of 4 to exaggerate differences in the upper range of performance and for between-group parity in expected earnings across payoff conditions. Both conditions were equally weighted across all rounds and used noisy but unscaled observations to assign a bonus of up to \$1.50. Subjects were informed in dollars about the bonus earned at the end of each round.

Environments. In Experiments 1 and 2, we used two classes of generated environments corresponding to different levels of smoothness (i.e., spatial correlation of rewards). These environments were sampled from a GP prior with a RBF kernel, where the length-scale parameter (λ) determines the rate at which the correlations of rewards decay over distance. *Rough* environments used $\lambda_{Rough} = 1$ and *Smooth* environments used $\lambda_{Smooth} = 2$, with 40 environments (Exp. 1) and 20 environments (Exp. 2) generated for each class (Smooth and Rough). In Experiment 3, we used environments defined by 20 real-world agricultural datasets, where the location on the grid corresponds to the rows and columns of a field and the payoffs reflect the normalized yield of various croups (see SI for full details).

Search horizons. We chose two horizon lengths (Short=5 or 20 and Long=10 or 40) that were fewer than the total number of tiles on the grid (30 or 121), and varied them within subject (alternating between

rounds and counterbalanced). Horizon length was approximately equivalent between Experiments 1 and Experiments 2 and 3, as a fraction of the total number of options (short $\approx \frac{1}{6}$; long $\approx \frac{1}{3}$).

Localization of Models

To penalize search options by the distance from the previous choice, we weighted each option by the inverse Manhattan distance (IMD) to the last revealed tile $\mathrm{IMD}(\mathbf{x}, \mathbf{x}') = (\sum_{i=1}^{n} |x_i - x_i'|)^{-1}$, prior to the softmax transformation. For the special case where $\mathbf{x} = \mathbf{x}'$, we set $\mathrm{IMD}(\mathbf{x}, \mathbf{x}') = 1$. Localized models are indicated by an asterix (*).

Model Comparison

We performed model comparison using cross-validated maximum likelihood estimation (MLE), where each participant's data was separated by horizon length (short or long) and we iteratively form a training set by leaving out a single round, compute a MLE on the training set, and then generate out-of-sample predictions on the remaining round (see SI for further details). This was repeated for all combinations of training set and test set, and for both short and long horizons. The cross-validation procedure yielded one set of parameter estimates per round, per participant, and out-of-sample predictions for 120 choices in Experiment 1 and 240 choices in Experiments 2 and 3 (per participant). Prediction error (computed as log loss) was summed up over all rounds, and is reported as *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:

$$R^2 = 1 - \frac{\log \mathcal{L}(\mathcal{M}_k)}{\log \mathcal{L}(\mathcal{M}_{\text{rand}})}, \tag{5}$$

where $\log \mathcal{L}(\mathcal{M}_{\text{rand}})$ is the log loss of a random model and $\log \mathcal{L}(\mathcal{M}_k)$ is model k's out-of-sample prediction error.

References

1. Todd, P. M., Hills, T. T. & Robbins, T. W. *Cognitive search: Evolution, algorithms, and the brain* (MIT press, 2012).
2. Kolling, N., Behrens, T. E., Mars, R. B. & Rushworth, M. F. Neural mechanisms of foraging. *Sci.* **336**, 95–98 (2012).
3. Bramley, N. R., Dayan, P., Griffiths, T. L. & Lagnado, D. A. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychol. Rev.* **124**, 301–338 (2017).
4. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press Cambridge, 1998).
5. Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. A Bayesian analysis of human decision-making on bandit problems. *J. Math. Psychol.* **53**, 168–179 (2009).
6. Speekenbrink, M. & Konstantinidis, E. Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).
7. Palminteri, S., Lefebvre, G., Kilford, E. J. & Blakemore, S.-J. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology* **13**, e1005684 (2017).
8. Reverdy, P. B., Srivastava, V. & Leonard, N. E. Modeling human decision making in generalized gaussian multiarmed bandits. *Proc. IEEE* **102**, 544–571 (2014).

9. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699 (2014).

10. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).

11. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40** (2017).

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

13. Tesauro, G. Practical issues in temporal difference learning. *Mach. learning* **8**, 257–277 (1992).

14. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nat.* **518**, 529–533 (2015).

15. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nat.* **529**, 484–489 (2016).

16. Huys, Q. J. *et al.* Interplay of approximate planning strategies. *Proc. Natl. Acad. Sci.* **112**, 3098–3103 (2015).

17. Solway, A. & Botvinick, M. M. Evidence integration in model-based tree search. *Proc. Natl. Acad. Sci.* **112**, 11708–11713 (2015).

18. Guez, A., Silver, D. & Dayan, P. Scalable and efficient Bayes-adaptive reinforcement learning based on monte-carlo tree search. *J. Artif. Intell. Res.* **48**, 841–883 (2013).

19. Rasmussen, C. E. & Kuss, M. Gaussian processes in reinforcement learning. In *Advances in Neural Information Processing Systems 16*, 751–758 (2004).

20. Sutton, R. S. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems 8*, 1038–1044 (1996).

21. Reverdy, P. B. & Leonard, N. E. Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Autom. Sci. Eng.* **13**, 54–67 (2016).

22. Lucas, C. G., Griffiths, T. L., Williams, J. J. & Kalish, M. L. A rational model of function learning. *Psychon. Bull. & Rev.* **22**, 1193–1215 (2015).

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

24. Borji, A. & Itti, L. Bayesian optimization explains human active search. In *Advances in Neural Information Processing Systems 26*, 55–63 (2013).

25. Dayan, P. & Niv, Y. Reinforcement learning: The good, the bad and the ugly. *Curr. opinion neurobiology* **18**, 185–196 (2008).

26. Srivastava, V., Reverdy, P. & Leonard, N. E. Correlated multiarmed bandit problem: Bayesian algorithms and regret analysis. *arXiv preprint arXiv:1507.01160* (2015).

27. Wilke, A. *et al.* A game of hide and seek: Expectations of clumpy resources influence hiding and searching patterns. *PloS One* **10**, e0130976 (2015).

28. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. Organizing conceptual knowledge in humans with a gridlike code. *Sci.* **352**, 1464–1468 (2016).

29. Stojic, H., Analytis, P. P. & Speekenbrink, M. Human behavior in contextual multi-armed bandit problems. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2290–2295 (2015).

30. Schulz, E., Konstantinidis, E. & Speekenbrink, M. Putting bandits into context: How function learning supports decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* (2017).

31. Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Connecting conceptual and spatial search via a model of generalization. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (2018). Forthcoming.

32. Hills, T. T., Jones, M. N. & Todd, P. M. Optimal foraging in semantic memory. *Psychol. Rev.* **119**, 431–440 (2012).

33. Abbott, J. T., Austerweil, J. L. & Griffiths, T. L. Random walks on semantic networks can resemble optimal foraging. *Psychol. Rev.* **122**, 558–569 (2015).

34. Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M. & Gershman, S. Assessing the perceived predictability of functions. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2116–2121 (2015).

35. Wright, K. *agridat: Agricultural Datasets* (2017). URL https://CRAN.R-project.org/package=agridat. R package version 1.13.

36. Lindley, D. V. On a measure of the information provided by an experiment. *The Annals Math. Stat.* 986–1005 (1956).

37. Nelson, J. D. Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychol. Rev.* **112** (2005).

38. Crupi, V. & Tentori, K. State of the field: Measuring information and confirmation. *Stud. Hist. Philos. Sci. Part A* **47**, 81–90 (2014).

39. Rasmussen, C. & Williams, C. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning (MIT Press, 2006).

40. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on gaussian process regression with a focus on exploration-exploitation scenarios. *bioRxiv* (2017).

41. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2002).

42. Neal, R. M. *Bayesian learning for neural networks* (Springer Science & Business Media, 1996).

43. Shepard, R. N. Toward a universal law of generalization for psychological science. *Sci.* **237**, 1317–1323 (1987).

44. Kaufmann, E., Cappé, O. & Garivier, A. On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, 592–600 (2012).

45. Palminteri, S., Wyart, V. & Koechlin, E. The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).

46. Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nat.* **441**, 876–879 (2006).

47. Metzen, J. H. Minimum regret search for single-and multi-task optimization. *arXiv preprint arXiv:1602.01064* (2016).

48. Gotovos, A., Casati, N., Hitz, G. & Krause, A. Active learning for level set estimation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1344–1350 (2013).

49. Cully, A., Clune, J., Tarapore, D. & Mouret, J.-B. Robots that can adapt like animals. *Nat.* **521**, 503–507 (2015).

50. Deisenroth, M. P., Fox, D. & Rasmussen, C. E. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis Mach. Intell.* **37**, 408–423 (2015).

51. Sui, Y., Gotovos, A., Burdick, J. & Krause, A. Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning*, 997–1005 (2015).

52. Srinivas, N., Krause, A., Kakade, S. & Seeger, M. W. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, 1015–1022 (2010).

53. Mockus, J. *Bayesian approach to global optimization: Theory and applications*, vol. 37 (Springer Science & Business Media, 2012).

54. Reece, S. & Roberts, S. An introduction to Gaussian processes for the Kalman filter expert. In *13th Conference on Information Fusion (FUSION)*, 1–9 (IEEE, 2010).

55. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat.* **521**, 436–444 (2015).

56. Schölkopf, B. Artificial intelligence: Learning to see and act. *Nat.* **518**, 486–487 (2015).

57. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).

Acknowledgments

We thank Peter Todd, Tim Pleskac, Neil Bramley, Henrik Singmann, and Mehdi Moussaïd for helpful feedback. This work was supported by the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World (CMW), by the Harvard Data Science Initiative (ES), and DFG grants ME 3717/2-2 to BM and NE 1713/1-2 to JDN as part of the New Frameworks of Rationality (SPP 1516) priority program. Code and data available at https://github.com/charleywu/gridsearch

Author Contributions

C.M.W. and E.S. designed the experiments, collected and analysed the data, and wrote the paper. M.S., J.D.N., and B.M. designed the experiments and wrote the paper.

Additional information

Competing financial interests

The author(s) declare no competing financial interests.