nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0467-4

Generalization guides human exploration in vast decision spaces

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

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 various bandit tasks with up to 121 arms, we study how humans search for rewards under limited search horizons, in which the spatial correlation of rewards (in both generated and natural environments) provides traction for generalization. Across various 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 insights about human behaviour in complex environments.

Many aspects of human behaviour can be understood as a type of search problem[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 problems[5–8], which imagine a gambler in front of a row of slot machines, learning the reward distributions of each option independently. 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). Yet, under real-world constraints of limited time or resources, it is not enough to know when to explore; one must also know where to explore.

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, in which 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 in which 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 efficient exploration when not all options can be explored?

One method for dealing with vast state spaces is to use function learning as a mechanism for generalizing previous 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. In addition, 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]. Although 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 (for example, in a search task, but see ref. [8]). Building on previous work exploring inductive biases in pure function learning contexts[21,22] and human behaviour in univariate function optimization[23], we present a comprehensive approach using a robust computational modelling framework to understand how humans generalize in an active search task.

Across three studies using univariate 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 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 about the nature of human generalization, showing the limits of traditional models of associative learning[24] in tasks in which the environmental structure supports learning and inference.

The main contributions of this paper are threefold:

(1) We introduce the spatially correlated multi-armed bandit as a paradigm for studying how people use generalization to guide search in larger problem spaces than traditionally used for studying human behaviour.
(2) We find that a Gaussian process 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.

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. [2]Department of Psychology, Harvard University, Cambridge, MA, 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. *e-mail: cwu@mpib-berlin.mpg.de

(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 a spatial correlation between rewards[25] (that is, clumpiness of resource distributions[26]). 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 not-yet-observed options, thereby guiding search decisions. In addition, as recent work has connected both spatial and conceptual representations to a common neural substrate[27], our results in a spatial domain provide potential pathways to other search domains, such as contextual[28–30] or semantic search[31,32].

Experiment 1. Participants ($n=81$) searched for rewards on a 1 × 30 grid world, in which 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. In addition, the search horizons (that is, number of clicks) alternated between rounds (within subject; short = 5 versus long = 10), with the order counterbalanced 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[33].

Looking first at sampling behaviour, the overall distance between sequential choices was more localized than chance ($t(80)=39.8$, $P<0.001$, $d=4.4$, 95% CI: 3.7–5.1, Bayes factor (BF) > 100; Fig. 1c; all reported t-tests are two sided), as has also been observed in semantic search[31] and causal learning[3] domains. Participants in the accumulation condition sampled more locally than those in the maximization condition ($t(79)=3.33$, $P=0.001$, $d=0.75$, 95% CI: 0.3–1.2, BF = 24), 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 ($t(79)=3.58$, $P<0.001$, $d=0.8$, 95% CI: 0.3–1.3, BF = 47.4), consistent with the hypothesis that spatial patterns in the environment can be learned and used to guide search. Surprisingly, longer search horizons (solid versus dashed lines in Fig. 1d) did not lead to higher average reward ($t(80)=0.60$, $P=0.549$, $d=0.07$, 95% CI: −0.4 to 0.5, BF = 0.2). We analysed both average reward and the maximum reward obtained for each subject, irrespective of their pay-off 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 ($t(79)=2.89$, $P=0.005$, $d=0.7$, 95% CI: 0.2–1.1, BF = 7.9), and performing equally well in terms of finding the largest overall reward ($t(79)=-0.73$, $P=0.467$, $d=-0.2$, 95% CI: −0.3 to 0.6, BF = 0.3). 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, middle panel) and longer search horizons to match the larger search space (short = 20 versus long = 40). We replicated the main results of experiment 1, showing that participants ($n=80$) sampled more locally than a random baseline ($t(79)=50.1$, $P<0.001$, $d=5.6$, 95% CI: 4.7–6.5, BF > 100; Fig. 1c), accumulation participants sampled more locally than maximization participants $t(78)=2.75$, $P=0.007$, $d=0.6$, 95% CI: 0.2–1.1, BF = 5.7), and participants obtained higher rewards in smooth than in rough environments ($t(78)=6.55$, $P<0.001$, $d=1.5$, 95% CI: 0.9–2.0, BF > 100; 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 largest reward values ($t(78)=-0.62$, $P=0.534$, $d=-0.1$, 95% CI: −0.6 to 0.3, BF = 0.3); yet, in experiment 2, the accumulation condition did not lead to substantially better performance than the maximization condition in terms of average reward ($t(78)=-1.31$, $P=0.192$, $d=-0.3$, 95% CI: −0.7 to 0.2, BF = 0.5). Again, short search horizons led to the same level of performance as long horizons, ($t(79)=-0.96$, $P=0.341$, $d=-0.1$, 95% CI: −0.3–0.1, BF = 0.2), 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 data sets[34], in which pay-offs correspond to the normalized yield of various crops (for example, wheat, corn and barley). These data sets have naturally occurring spatial correlations and are naturally segmented into a grid based on the rows and columns of a field, thus requiring no interpolation or other transformation except for the normalization of pay-offs (see Supplementary Information for selection criteria). The crucial difference compared to experiment 2 is that these natural data sets comprise a set of more complex environments in which learners could nonetheless still benefit from spatial generalization.

As in both previous experiments, participants ($n=80$) sampled more locally than random chance ($t(79)=50.1$, $P<0.001$, $d=5.6$, 95% CI: 4.7–6.5, BF > 100), with participants in the accumulation condition sampling more locally than those in the maximization condition ($t(78)=3.1$, $P=0.003$, $d=0.7$, 95% CI: 0.2–1.1, BF = 12.1). In the natural environments, we found that accumulation participants achieved a higher average reward than maximization participants ($t(78)=2.7$, $P=0.008$, $d=0.6$, 95% CI: 0.2–1.1, BF = 5.6), with an effect size similar to experiment 1. There was no difference in maximum reward across pay-off conditions ($t(78)=0.3$, $P=0.8$, $d=0.06$, 95% CI: −0.4 to 0.5, BF = 0.2), as in all previous experiments, showing that the goal of balancing exploration–exploitation leads to the best results on both performance metrics. As in the previous experiments, we found that a longer search horizon did not lead to higher average rewards ($t(78)=2.1$, $P=0.04$, $d=0.2$, 95% CI: −0.2 to 0.7, BF = 0.4). Thus, the results of experiment 3 closely corroborate the results of experiments 1 and 2, showing that our findings on human behaviour in simulated environments are very similar to human behaviour in natural environments.

Modelling generalization and search. To better understand how participants explore, we compared a diverse set of computational models in their ability to predict each subject's trial-by-trial choices (see Supplementary Fig. 1 and Supplementary Table 3 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 that is central to theories in statistics[35], psychology[36], and philosophy of science[37]. Models of learning

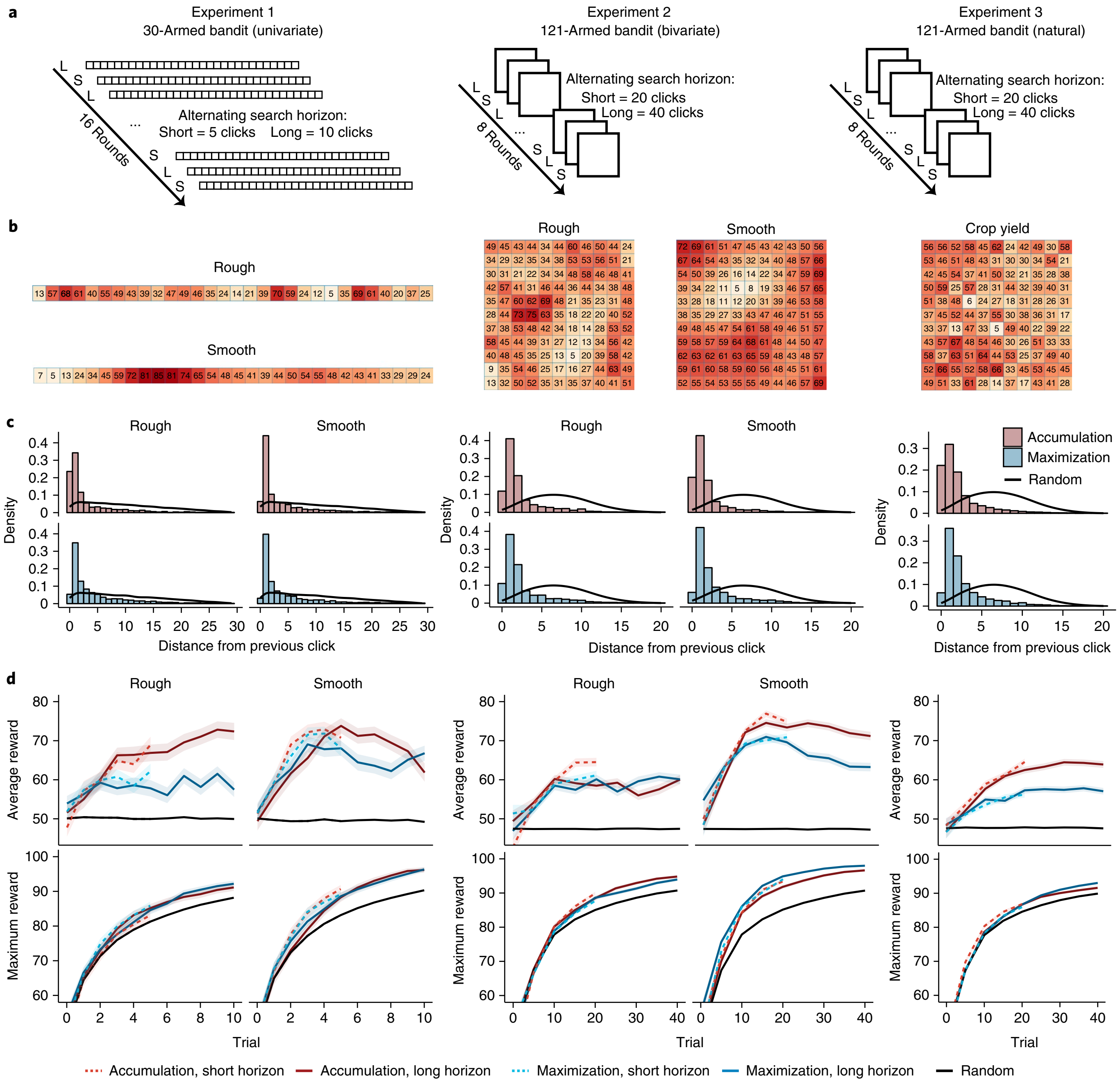

Fig. 1 | Procedure and behavioural results. Experiments 1 and 2 used a 2 × 2 between-subject design, manipulating the type of environment (rough or smooth) and the pay-off condition (accumulation or maximization), whereas experiment 3 manipulated only pay-off conditions (between subjects) and used a set of natural environments where rewards reflect normalized crop yields from various agricultural data sets. **a**, Experiment 1 used a 1D array of 30 possible options, whereas experiments 2 and 3 used a 2D array (11 × 11) with 121 options. Experiments took place over 16 (experiment 1) or 8 (experiments 2 and 3) rounds, with a new environment sampled without replacement for each round. Search horizons alternated between rounds (within subject), with the horizon order counterbalanced between subjects. L, long; S, short. **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, whereas crop yield environments have no fixed level of correlation (see Supplementary Information). **c**, Locality of sampling behaviour compared with a random baseline simulated over 10,000 rounds (black line), in which distance is measured using Manhattan distance and the *y* axis indicates the probability density of different distances (with a different maximum range for experiment 1 compared to experiments 2 and 3). **d**, Average reward earned (accumulation goal) and maximum reward revealed (maximization goal), in which the coloured lines indicate the assigned pay-off condition and the shaded regions show the standard error of the mean. Black lines indicate a random baseline simulated over 10,000 rounds.

form inductive beliefs about the value of possible options (including unobserved options) conditioned on previous observations, whereas sampling strategies transform these beliefs into probabilistic predictions about where a participant will sample next. We also consider heuristics, which are competitive models of human behaviour in bandit tasks[5], yet do not maintain a model of the world (see Supplementary Information). By far the best-predictive models used Gaussian process regression[38,39] as a mechanism for generalization

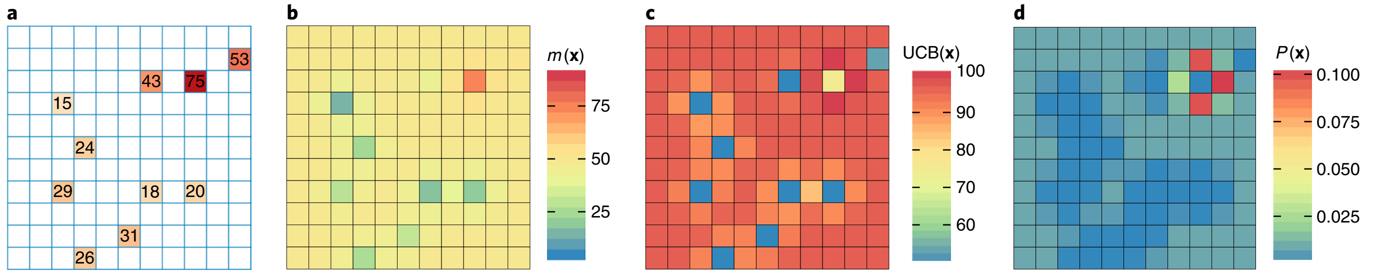

Fig. 2 | Overview of the function learning–UCB model specified using median participant parameter estimates from experiment 2. a, Screenshot of experiment 2. Participants were allowed to select any tile until the search horizon was exhausted. **b**, Estimated reward (the estimated uncertainty is not shown) as predicted by the Gaussian process function learning model, based on the points sampled in **a**. **c**, UCB of predicted rewards. **d**, Choice probabilities after a softmax choice rule. $P(\mathrm{x}) = \exp(\mathrm{UCB}(\mathrm{x})/\tau)/\sum_{j=1}^{N}\exp(\mathrm{UCB}(\mathrm{x}_j)/\tau)$, where τ is the temperature parameter (that is, higher temperature values lead to more undirected, noisy sampling). For parameter estimates, see Supplementary Table 3.

and UCB sampling[40] 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 Gaussian process regression as an expressive model of human function learning, which has known equivalencies to neural network function approximators[41], yet provides psychologically interpretable parameter estimates about the extent to which generalization occurs. Gaussian process 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. 2a,b), conditioned on a finite number of previous observations of rewards $\mathbf{y}_T = [y_1, y_2, \ldots, y_T]^{\mathrm{T}}$ 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):

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

where the length-scale parameter λ governs how quickly correlations between points $\mathbf{x}$ and $\mathbf{x}'$ (for example, 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. As the Gaussian process prior is completely defined by the RBF kernel, the underlying mechanisms are similar to Shepard's universal gradient of generalization[42], 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 (effectively) 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:

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

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 that participants would prefer a hypothetical option $\mathbf{x}_1$ predicted to have mean reward $m(\mathbf{x}_1) = 60$ and standard deviation $s(\mathbf{x}_1) = 10$, over an option $\mathbf{x}_2$ predicted to have mean reward $m(\mathbf{x}_2) = 64$ and standard deviation $s(\mathbf{x}_2) = 1$. This is because sampling $\mathbf{x}_1$ is expected to reduce a large amount of uncertainty, even though $\mathbf{x}_2$ has a higher mean reward (as $\mathrm{UCB}(\mathbf{x}_1) = 65$ but $\mathrm{UCB}(\mathbf{x}_2) = 64.5$). This trade-off between exploiting known high-value options and exploring to reduce uncertainty[43] 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 based on expected reward (pure exploitation) or uncertainty (pure exploration):

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

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

Figure 2 shows how the Gaussian process–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 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 in which 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 (that is, higher temperatures correspond to more noisy sampling; Fig. 2d), in contrast to the β parameter of UCB, which estimates the level of exploration directed towards reducing uncertainty.

Modelling results

Experiment 1. Participants were better described by the function learning model than the option learning model ($t(80) = 14.10$, $P < 0.001$ $d = 1.6$, 95% CI: 1.1–2.1, $BF > 100$, 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 the pure exploitation ($t(80) = -8.85$, $P < 0.001$, $d = -1.0$, 95% CI: −0.5 to −1.4), $\mathrm{BF} > 100$) and pure exploration ($t(80) = -16.63$, $P < 0.001$, $d = -1.8$,

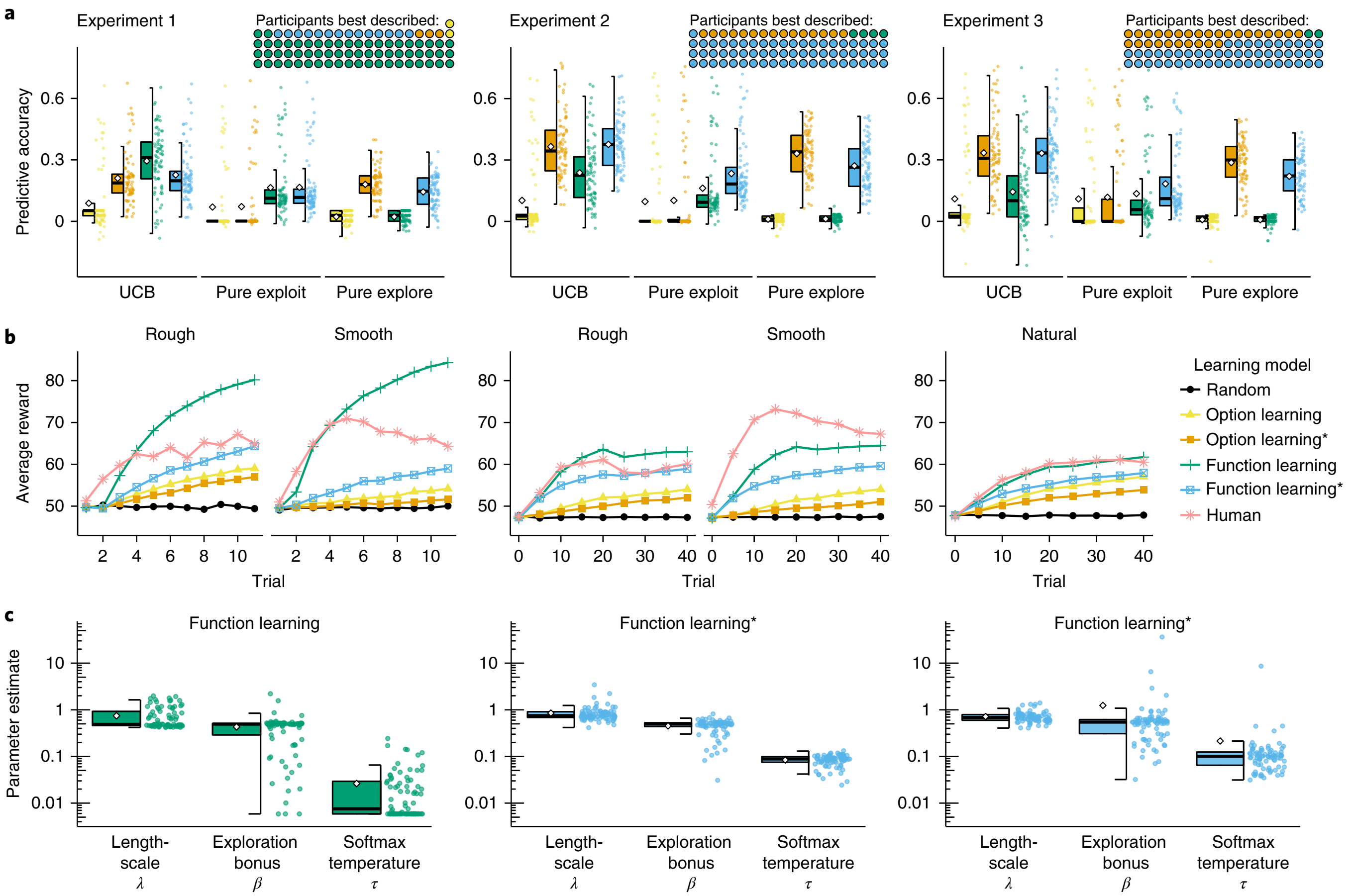

Fig. 3 | Modelling results. **a**, Cross-validated predictive accuracy of each model (higher is better), with box plots indicating the interquartile range (box), the median (horizontal line), mean (diamond) and 1.5-times interquartile range (whiskers). Each individual participant is shown as a single dot, with the number of participants best described shown as an icon array (inset; aggregated by sampling strategies). Colours indicate the learning model (see panel **b** caption) where asterisks (*) indicate a localized variant of the option learning or function learning models, in which predictions are weighted by the inverse distance from the previous choice (see Methods). **b**, Learning curves of participants and model simulations. Each simulated learning model uses UCB sampling and is specified using participant parameter estimates and averaged over 100 simulated experiments per participant per model. **c**, Parameter estimates of the best-predicting model for each experiment. Each coloured dot is the median estimate per participant, with box plots indicating in the interquartile range (box), 1.5-times interquartile range (whiskers), median (horizontal line) and mean (diamond).

95% CI: −1.3 to −2.4, BF > 100) variants each made less accurate 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). Although the option learning* model was better than the standard option learning model ($t(80) = 16.13$, $P < 0.001$, $d = 1.8$, 95% CI: 1.3–2.3, BF > 100), the standard function learning model still outperformed its localized variant ($t(80) = 5.05$, $P < 0.001$, $d = 0.6$, 95% CI: 0.1–1.0, BF > 100). 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. Finally, we also calculated each model's protected probability of exceedance[44] using its out-of-sample log-evidence. This probability assesses which model is the most common among all models in our pool (among the 12 models reported in the main text; see Supplementary Table 3 for a comparison with additional models) while also correcting for chance. Doing so, we found that the function learning–UCB model reached a protected probability of pxp = 1, indicating that it vastly outperformed all of the other models.

Figure 3b shows simulated learning curves of each model in comparison to human performance, in which models were specified using parameters from participants' estimates (curves averaged over 100 simulated experiments per participant per model). Whereas both versions of the option learning model improve only very slowly, both standard and localized versions of the function learning model behave sensibly and show a close alignment to the rapid rate of human learning during the early phases of learning. However, there is still a deviation in similarity between the curves, which is partially due to aggregating over reward conditions and horizon manipulations, in addition to aggregating over individuals, where some participants over-explore their environments, whereas others produce continuously increasing learning curves (see Supplementary Fig. 6 for individual learning curves). Although aggregated learning curves should be analysed with caution[45], we find 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.

We compare participants' parameter estimates using a Wilcoxon signed rank test to make the resulting differences more robust to potential outliers. The parameter estimates of the function learning model (Fig. 3c) indicated that people tend to underestimate the extent of spatial correlations, with median per-participant λ estimates significantly lower than the ground truth ($\lambda_{\mathrm{Smooth}} = 2$ and $\lambda_{\mathrm{Rough}} = 1$) for both smooth environments (Wilcoxon signed rank test; $\hat{\lambda}_{Smooth} = 0.5$, $Z = -7.1$, $P < 0.001$, $r = 1.1$, $\mathrm{BF}_Z > 100$) and rough environments ($\hat{\lambda}_{Rough} = 0.5$, $Z = -3.4$, $P < 0.001$, $r = 0.55$, $\mathrm{BF}_Z > 100$). This can be interpreted as a tendency towards undergeneralization. In addition, we found that the estimated exploration bonus of UCB sampling (β) was reliably greater than zero ($\hat{\beta} = 0.51$, $Z = -7.7$, $P < 0.001$, $r = 0.86$, $\mathrm{BF}_Z > 100$, than the lower estimation bound), reflecting the valuation of sampling uncertain options, together with exploiting high expectations of reward. Finally, 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 the 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 ($t(79) = 9.99$, $P < 0.001$, $d = 1.1$, 95% CI: 0.6–1.6, BF > 100), although this was only marginally the case when comparing localized function learning* to localized option learning* ($t(79) = 2.05$, $P = 0.044$, $d = 0.2$, 95% CI: −0.2 to 0.7, BF = 0.9). In the two-dimensional (2D) search environment of experiment 2, adding localization improved predictions for both option learning ($t(79) = 19.92$, $P < 0.001$, $d = 2.2$, 95% CI: 1.7–2.8, BF > 100) and function learning ($t(79) = 10.47$, $P < 0.001$, $d = 1.2$, 95% CI: 0.7–1.6, BF > 100), in line with the stronger tendency towards localized sampling than 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 predicted by the localized option learning* model. Again, both components of the UCB strategy were necessary to predict choices, with pure exploit ($t(79) = -6.44$, $P < 0.001$, $d = -0.7$, 95% CI: −0.3 to −1.2, BF > 100) and pure explore ($t(79) = -12.8$, $P < 0.001$, $d = -1.4$, 95% CI: −0.9 to −1.9, BF > 100) making worse predictions. The probability of exceedance over all models showed that the function learning*–UCB model achieved virtually pxp = 1, indicating that it greatly outperformed all other models under consideration.

As in experiment 1, the simulated learning curves of the option learning models increased slowly and only marginally outperformed 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, although participants generalized somewhat more than in experiment 1 ($\hat{\lambda} = 0.75$, $Z = -3.7$, $P < 0.001$, $r = 0.29$, $\mathrm{BF}_Z > 100$), they again underestimated the strength of the underlying spatial correlation in both smooth environments ($\hat{\lambda}_{Smooth} = 0.78$, $Z = -5.8$, $P < 0.001$, $r = 0.88$, $\mathrm{BF}_Z > 100$; comparison to $\lambda_{\mathrm{Smooth}} = 2$) and rough environments ($\hat{\lambda}_{Rough} = 0.75$, $Z = -4.7$, $P < 0.001$, $r = 0.78$, $\mathrm{BF}_Z > 100$; comparison to $\lambda_{\mathrm{Rough}} = 1$). This suggests a robust tendency to undergeneralize. There were no differences in the estimated exploration bonus β between experiments 1 and 2 ($\hat{\beta} = 0.5$, $Z = 0.86$, $P = 0.80$, $r = 0.07$, $\mathrm{BF}_Z = 0.2$), although the estimated softmax temperature parameter τ was larger than in experiment 1 ($\hat{\tau} = 0.09$, $Z = -8.89$, $P < 0.001$, $r = 0.70$, $\mathrm{BF}_Z = 34$). Thus, experiment 2 replicated the main findings of experiment 1. When taken together, results from the two experiments provide strong evidence that human search behaviour is best explained by 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 replicated key results from the previous experiments: function learning made better predictions than option learning ($t(79) = 3.03$, $P = 0.003$, $d = 0.3$, 95% CI: −0.1 to 0.8, BF = 8.2); adding localization improved predictions for both option learning ($t(79) = 18.83$, $P < 0.001$, $d = 2.1$, 95% CI: 1.6–2.6, BF > 100) and function learning ($t(79) = 14.61$, $P < 0.001$, $d = 1.6$, 95% CI: 1.1–2.1, BF > 100); and the combined UCB algorithm performed better than using only a pure exploit strategy ($t(79) = 12.97$, $P < 0.001$, $d = 1.4$, 95% CI: 1.0–1.9, BF > 100) or a pure explore strategy ($t(79) = 5.87$, $P < 0.001$, $d = 0.7$, 95% CI: 0.3–1.2, BF > 100). However, the difference between the localized function learning* and the localized option learning* was negligible ($t(79) = 0.32$, $P = 0.75$, $d = 0.04$, 95% CI: −0.4 to 0.5, BF = 0.1). This is perhaps owing to the high variability across environments, which makes it harder to predict out-of-sample choices using generalization behaviour (that is, λ) estimated from a separate set of environments. Nevertheless, the localized function learning* model was still the best-predicting model for the majority of participants (48 out of 80 participants). Moreover, calculating the protected probability of exceedance over all models' predictive evidence revealed a probability of pxp = 0.98 that the function learning* model was more frequent in the population than all of the other models, followed by pxp = 0.01 for the option learning* model. Thus, even in natural environments in which the underlying spatial correlations are unknown, we were still able to distinguish the different models in terms of their overall out-of-sample predictive performance.

The simulated learning curves in Fig. 3b show the strongest concurrence out of all previous experiments between the function learning model and human performance. Moreover, both variants of the option learning model learn far slower, failing to match the rate of human learning, suggesting that they are not plausible models of human behaviour[46]. 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 ($\hat{\lambda}_{natural} = 0.68$, $Z = -3.4$, $P < 0.001$, $r = 0.27$, $\mathrm{BF}_Z = 9.6$) and exploring slightly more, with a small increase in both directed exploration ($\hat{\beta}_{natural} = 0.54$, $Z = -2.3$, $P = 0.01$, $r = 0.18$, $\mathrm{BF}_Z = 4.5$) and undirected exploration ($\hat{\tau}_{natural} = 0.1$, $Z = -2.2$, $P = 0.02$, $r = 0.17$, $\mathrm{BF}_Z = 4.2$) 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 Supplementary Information). Model recovery consisted of simulating data using a generating model specified by participant parameter estimates. We then performed the same cross-validation procedure to fit a recovering model on this simulated data. In all cases, the best-predictive accuracy occurred when the recovering model matched the generating model (Supplementary Fig. 2), suggesting robustness to type I errors and ruling out model overfitting (that is, 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 (Supplementary Fig. 3). 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 phenomenon from noisy, undirected exploration[47].

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. Thus, 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. Remarkably, undergeneralization sometimes is even better than exactly matching the underlying

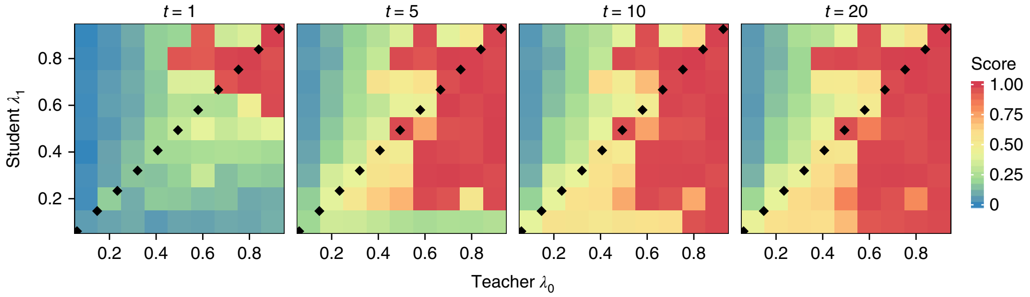

Fig. 4 | Mismatched length-scale (λ) simulation results. Each panel is performance at a different trial t. The teacher λ_0 values were used to generate environments, whereas the student λ_1 values were used to parameterize the function learning model to simulate search performance. The dotted lines show where $\lambda_0 = \lambda_1$ and mark the difference between undergeneralization and overgeneralization, with points below the diagonal 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.

structure of the environment (Fig. 4). These simulations were performed by first generating search environments by sampling from a Gaussian process 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[48] with continuous bivariate inputs in the range x, $y = [0,1]$, allowing for a broader set of potential mismatched alignments (see Supplementary Fig. 4 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 (that is, 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 are 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 (Supplementary Fig. 4c). 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 Gaussian process regression, provides a mechanism for generalization, which can be used to guide search towards unexplored yet promising options. Combined with UCB sampling, this model navigates the exploration–exploitation dilemma by optimistically inflating expectations of reward by the estimated uncertainty.

Although Gaussian process function learning combined with a UCB sampling algorithm has been successfully applied to search problems in ecology[49], robotics[50,51] and biology[52], 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 importance, and our work makes key theoretical and empirical contributions. Expanding on previous studies that found an overlap between Gaussian process–UCB and human learning rates[8,23], we 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 that 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 (Supplementary Fig. 2) and parameter estimates (Supplementary Fig. 3). Whereas previous research on exploration bonuses has had mixed results[6,12,47], 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[53]. This shows a remarkable concurrence between intuitive human strategies and state-of-the-art machine learning research.

Limitations and extensions. One potential limitation is that our pay-off manipulation (maximization versus accumulation) failed to induce superior performance according to the relevant performance metric. Although participants in the accumulation condition achieved higher average reward, participants in the maximization condition were not able to outperform with respect to the maximum reward criterion. The goal of balancing exploration–exploitation (accumulation condition) or the goal of global optimization (maximization condition) was induced through the manipulation of written instructions, comprehension check questions and feedback between rounds (see Methods). Although this may have been insufficient for observing clear performance differences (but see Supplementary Table 1), the practical difference between these two goals is murky even in the Bayesian optimization literature, in which the strict goal of finding the global optimum is often abandoned based purely on computational concerns[54]. Instead, the global optimization goal is frequently replaced by an approximate measure of performance, such as cumulative regret[53], which closely aligns to our accumulation pay-off condition. In our experiments, remarkably, participants assigned to the accumulation goal pay-off condition also performed best relative to the maximization criterion.

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 a special case of Gaussian process regression[55]. 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 independence to the opposite extreme of complete state interdependence. Moreover, Gaussian processes also have equivalencies to Bayesian neural networks[41], suggesting a further link to distributed function learning models[56]. Indeed, one explanation for the impressive performance of deep reinforcement learning[14] is that neural networks are specifically a powerful type of function approximator[57].

Finally, both spatial and conceptual representations have been connected to a common neural substrate in the hippocampus[27], suggesting a potential avenue for applying the same function learning–UCB model for modelling human learning using contextual[28–30], semantic[31,32] or potentially even graph-based features. One hypothesis for this common role of the hippocampus is that it performs predictive coding of future state transitions[58], also known as 'successor representation'[24]. In our task, in which there are no restrictions on state transitions (that is, each state is reachable from any previous state), it may be the case that the RBF kernel driving our Gaussian process function learning model performs the same role as the transition matrix of a successor representation model, in which state transitions are learned via a random walk policy.

Conclusions

We present a paradigm for studying how people use generalization to guide the active search for rewards and found a systematic—yet sometimes beneficial—tendency to undergeneralize. In addition, we uncovered substantial 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 the 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. Participants ($n=81$) were recruited from Amazon Mechanical Turk for experiment 1 (25 female; mean ± s.d. age: 33 ± 11 years), 80 for experiment 2 (25 female; mean ± s.d. age: 32 ± 9 years) and 80 for experiment 3 (24 female; mean ± s.d. age: 35 ± 10 years). In all of the experiments, participants were paid a participation fee of US\$0.50 and a performance contingent bonus of up to US\$1.50. Participants earned on average US\$1.14 ± 0.13 and spent 8 ± 4 min on the task in experiment 1, earned US\$1.64 ± 0.20 and spent 8 ± 4 min in experiment 2, and earned US\$1.53 ± 0.15 and spent 8 ± 5 min in experiment 3. Participants were only allowed to participate in one of the experiments and were required to have a 95% human interaction task (HIT) approval rate and 1,000 previously completed HITs. No statistical methods were used to pre-determine sample sizes, but our sample sizes are similar or larger to those reported in previous publications[6,12,23,28,29]. 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, in which participants were randomly assigned to one of two different pay-off structures (accumulation condition versus maximization condition) and one of two different classes of environments (smooth versus rough), whereas experiment 3 used environments from real-world agricultural data sets and manipulated only the pay-off structure (random assignment between subjects). Each grid world represented a (either univariate or bivariate) function, with each observation including normally distributed noise, $\varepsilon \sim \mathcal{N}(0,1)$. The task was presented over either 16 rounds (experiment 1) or 8 rounds (experiments 2 and 3) on different grid worlds, which were randomly drawn (without replacement) from the same class of environments (that is, same length-scale parameter λ). Participants had either a short or long search horizon (short = 5 and long = 10 trials in experiment 1; short = 20 and long = 40 trials in experiments 2 and 3) to sample tiles on the grid, including repeat clicks. The search horizon alternated between rounds (within subject), with initial horizon length counterbalanced between subjects by random assignment. Data collection and analysis were not performed blind to the conditions of the experiments.

Materials and procedure. Before 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, in which darker colours indicated higher point values. Per round, observations were scaled to a randomly drawn maximum value in the range of 65–85, so that the value of the global optima could not be easily guessed (for example, 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, whereas hovering over the tile would display the entire history of observation. The colour of the tile corresponded to the mean of all previous observations.

Pay-off conditions. We compared performance under two different pay-off conditions, requiring either a balance between exploration and exploitation (accumulation condition) or corresponding to consistently making exploration decisions (maximization condition). In each pay-off condition, participants received a performance contingent bonus of up to US\$1.50. Accumulation condition participants were given a bonus based on the average value of all clicks as a fraction of the global optima, $\frac{1}{T}\sum\left(\frac{y_t}{y^*}\right)$, 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, $\left(\frac{\max y_t}{y^*}\right)^4$, taken to the power of 4 to exaggerate differences in the upper range of performance and for between-group parity in expected earnings across pay-off conditions. Both conditions were equally weighted across all rounds and used noisy but unscaled observations to assign a bonus of up to US\$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 (that is, spatial correlation of rewards). These environments were sampled from a Gaussian process prior with a RBF kernel, in which the length-scale parameter (λ) determines the rate at which the correlations of rewards decay over distance. Rough environments used $\lambda_{\text{Rough}}=1$ and smooth environments used $\lambda_{\text{Smooth}}=2$, with 40 environments (experiment 1) and 20 environments (experiment 2) generated for each class (smooth and rough). In experiment 3, we used environments defined by 20 real-world agricultural data sets, where the location on the grid corresponds to the rows and columns of a field and the rewards reflect the normalized yield of various crops (see Supplementary Information 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 experiment 1 and experiments 2 and 3, as a fraction of the total number of options $\left(\text{short}\approx\frac{1}{6};\text{long}\approx\frac{1}{3}\right)$.

Statistical tests. All reported t-tests are two sided. We also report Bayes factors (BF), quantifying the likelihood of the data under H_A relative to the likelihood of the data under H_0. We calculate the default two-sided Bayesian t-test using a Jeffreys–Zellner–Siow prior with its scale set to $\sqrt{2}/2$, following ref. [59]. For parametric tests, the data distribution was assumed to be normal, but this was not formally tested. For non-parametric comparisons, the Bayes factor BF_Z is derived by performing posterior inference over the Wilcoxon test statistics and assigning a prior by means of a parametric yoking procedure[60]. The null hypothesis posits that the statistic between two groups does not differ, and the alternative hypothesis posits the presence of an effect and assigns an effect size using a Cauchy distribution with the scale parameter set to $1/\sqrt{2}$.

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 $\text{IMD}(\mathbf{x},\mathbf{x}')=\left(\sum_{i=1}^{n}|x_i - x_i'|\right)^{-1}$, prior to the softmax transformation. For the special case where $\mathbf{x}=\mathbf{x}'$, we set $\text{IMD}(\mathbf{x},\mathbf{x}')=1$. Localized models are indicated by an asterisk (*).

Model comparison. We performed model comparison using cross-validated maximum likelihood estimation, in which each participant's data were separated by horizon length (short or long) and we iteratively formed a training set by leaving out a single round, compute a maximum likelihood estimation on the training set and then generate out-of-sample predictions on the remaining round (see Supplementary Information 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:

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

where $\log\mathcal{L}(\mathcal{M}_{\text{rand}})$ is the log loss of a random model and $\log\mathcal{L}(\mathcal{M}_k)$ is the model k's out-of-sample prediction error. Moreover, we calculated each model's protected probability of exceedance using its predictive log evidence[44]. This probability is defined as the probability that a particular model is more frequent in the population than all of the other models, averaged over the probability of the null hypothesis that all models are equally frequent (thereby correcting for chance performance).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The code used for all models and analyses is available at https://github.com/charleywu/gridsearch.

Data availability

Anonymized participant data and model simulation data are available at https://github.com/charleywu/gridsearch.

Received: 2 August 2017; Accepted: 4 October 2018;
Published online: 12 November 2018

References

1. Todd, P. M., Hills, T. T. & Robbins, T. W. *Cognitive Search: Evolution, Algorithms, and the Brain* (MIT Press, Cambridge, 2012).
2. Kolling, N., Behrens, T. E., Mars, R. B. & Rushworth, M. F. Neural mechanisms of foraging. *Science* **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 Comput. Biol.* **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**, e253 (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. Learn.* **8**, 257–277 (1992).
14. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
15. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
16. Huys, Q. J. et al. Interplay of approximate planning strategies. *Proc. Natl Acad. Sci. USA* **112**, 3098–3103 (2015).
17. Solway, A. & Botvinick, M. M. Evidence integration in model-based tree search. *Proc. Natl Acad. Sci. USA* **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. *Adv. Neural Inf. Process. Syst.* **16**, 751–758 (2004).
20. Sutton, R. S. Generalization in reinforcement learning: successful examples using sparse coarse coding. *Adv. Neural Inf. Process. Syst.* **8**, 1038–1044 (1996).
21. 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).
22. Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M. & Gershman, S. J. Compositional inductive biases in function learning. *Cogn. Psychol.* **99**, 44–79 (2017).
23. Borji, A. & Itti, L. Bayesian optimization explains human active search. *Adv. Neural Inf. Process. Syst.* **26**, 55–63 (2013).
24. Dayan, P. & Niv, Y. Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* **18**, 185–196 (2008).
25. Srivastava, V., Reverdy, P. & Leonard, N. E. Correlated multiarmed bandit problem: Bayesian algorithms and regret analysis. Preprint at https://arxiv.org/abs/1507.01160 (2015).
26. Wilke, A. et al. A game of hide and seek: expectations of clumpy resources influence hiding and searching patterns. *PLoS ONE* **10**, e0130976 (2015).
27. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
28. Stojic, H., Analytis, P. P. & Speekenbrink, M. Human behavior in contextual multi-armed bandit problems. In *Proc. 37th Annual Meeting of the Cognitive Science Society* (eds Noelle, D. C. et al.) 2290–2295 (Cognitive Science Society, 2015).
29. Schulz, E., Konstantinidis, E. & Speekenbrink, M. Putting bandits into context: how function learning supports decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 927–943 (2018).
30. Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Connecting conceptual and spatial search via a model of generalization. In *Proc. 40th Annual Meeting of the Cognitive Science Society* (eds Rogers, T. T., Rau, M., Zhu, X. & Kalish, C. W.) 1183–1188 (Cognitive Science Society, 2018).
31. Hills, T. T., Jones, M. N. & Todd, P. M. Optimal foraging in semantic memory. *Psychol. Rev.* **119**, 431–440 (2012).
32. Abbott, J. T., Austerweil, J. L. & Griffiths, T. L. Random walks on semantic networks can resemble optimal foraging. *Psychol. Rev.* **122**, 558–569 (2015).
33. Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M. & Gershman, S. Assessing the perceived predictability of functions. In *Proc. 37th Annual Meeting of the Cognitive Science Society* (eds Noelle, D. C. et al.) 2116–2121 (Cognitive Science Society, 2015).
34. Wright, K. agridat: Agricultural Datasets R Package Version 1.13 (2017); https://CRAN.R-project.org/package=agridat
35. Lindley, D. V. On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986–1005 (1956).
36. Nelson, J. D. Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychol. Rev.* **112**, 979–999 (2005).
37. Crupi, V. & Tentori, K. State of the field: measuring information and confirmation. *Stud. Hist. Philos. Sci. A* **47**, 81–90 (2014).
38. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (MIT Press, Cambridge, 2006).
39. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).
40. Auer, P. Using confidence bounds for exploitation–exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2002).
41. Neal, R. M. *Bayesian Learning for Neural Networks* (Springer, New York, 1996).
42. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).
43. Kaufmann, E., Cappé, O. & Garivier, A. On Bayesian upper confidence bounds for bandit problems. In *Proc. 15th International Conference on Artificial Intelligence and Statistics (AISTAT)* (eds Lawrence, N. D. & Girolami, M. A.) 592–600 (JMLR, 2012).
44. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017 (2009).
45. Myung, I. J., Kim, C. & Pitt, M. A. Toward an explanation of the power law artifact: insights from response surface analysis. *Mem. Cognit.* **28**, 832–840 (2000).
46. Palminteri, S., Wyart, V. & Koechlin, E. The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
47. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
48. Metzen, J. H. Minimum regret search for single- and multi-task optimization. Preprint at https://arxiv.org/abs/1602.01064 (2016).
49. Gotovos, A., Casati, N., Hitz, G. & Krause, A. Active learning for level set estimation. In *International Joint Conference on Artificial Intelligence (IJCAI)* (ed. Rossi, F.) 1344–1350 (AAAI Press/International Joint Conferences on Artificial Intelligence, 2013).
50. Cully, A., Clune, J., Tarapore, D. & Mouret, J.-B. Robots that can adapt like animals. *Nature* **521**, 503–507 (2015).

51. Deisenroth, M. P., Fox, D. & Rasmussen, C. E. Gaussian processes for data-efficient learning in robotics and control. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 408–423 (2015).
52. Sui, Y., Gotovos, A., Burdick, J. & Krause, A. Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning* (eds Bach, F. & Blei, D.) 997–1005 (PMLR, 2015).
53. Srinivas, N., Krause, A., Kakade, S. & Seeger, M. W. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proc. 27th International Conference on Machine Learning* (eds Fürnkranz, J. & Joachims, T.) 1015–1022 (Omnipress, 2010).
54. Mockus, J. *Bayesian Approach to Global Optimization: Theory and Applications* Vol. 37 (Springer, Dordrecht, 2012).
55. 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).
56. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
57. Schölkopf, B. Artificial intelligence: learning to see and act. *Nature* **518**, 486–487 (2015).
58. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
59. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).
60. van Doorn, J., Ly, A., Marsman, M. & Wagenmakers, E. J. Bayesian latent-normal inference for the rank sum test, the signed rank test, and Spearman's ρ. Preprint at https://arxiv.org/abs/1712.06941 (2017).

Acknowledgements

We thank P. Todd, T. Pleskac, N. Bramley, H. Singmann and M. Moussaïd for helpful feedback. This work was supported by the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World (C.M.W.), by the Harvard Data Science Initiative (E.S.), and DFG grants ME 3717/2-2 to B.M. and NE 1713/1-2 to J.D.N. as part of the New Frameworks of Rationality (SPP 1516) priority programme. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0467-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.M.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Charley M. Wu

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We used online experiments, where experiment code is freely available at https://github.com/charleywu/gridsearch
Data analysis	All code used to analyze the data is freely available at https://github.com/charleywu/gridsearch

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Experiment data is freely available online at https://github.com/charleywu/gridsearch

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The studies presented here were conducted online using Amazon Mechanical Turk (MTurk). Each experiment took the form of an online game, where performance was incentivized through a paid bonus. We collected quantitative data based on each individual's choice at each trial. Participants were first given instructions for the task along with several examples of fully revealed environments. Then participants were asked to first complete a set of comprehension questions before starting the experiment. Between rounds, participants were informed in dollar amounts (USD) how well they performed on the previous round. At the end of the task, participants were asked to provide demographic information and to confirm their MTurk worker id that was used to assign them their performance bonus.
Research sample	Participants were recruited on Amazon MTurk (see Recruitment section below for exclusion criteria). We recruited 81 participants for Experiment 1 (25 Female; mean age ± SD 33 ± 11), 80 for Experiment 2 (25 Female; mean age ± SD 32 ± 9), and 80 for Experiment 3 (24 Female; mean age ± SD 35 ± 10)
Sampling strategy	The sample size was pre-determined such that the 2x2 cross-over design would allow for about 40 participants in each between-group. Since the focus on the quantitative analysis was on computational modeling, rather than purely behavioral analysis, our sample size determination was not focused on achieving the necessary power to observe a specific effect size.
Data collection	All experiments were collected online, with experiment code available freely online at https://github.com/charleywu/gridsearch
Timing	Participants were given 1 hour to complete the experiment, before the HIT expired and would be re-published to another MTurk worker. This was well beyond the typically completion time (Experiment 1: 8±4 minutes; Experiment 2: 8±4 minutes; Experiment 3: 8±5 minutes)
Data exclusions	No collected data was excluded from any experiments.
Non-participation	We had no non-participants
Randomization	Participants were assigned experimental conditions based on a pre-randomized list.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	see above
Recruitment	Participants were recruited on Amazon MTurk, requiring a 95% HIT approval rate and 1000 previously completed HITs. Participants who completed one of the experiments in this paper were excluded from subsequent experiments by granting them a qualification, which was then used as an exclusion criteria.